Pseudo-Labeling Enhanced by Privileged Information and its application to In Situ Sequencing Images

IJCAI23
Supplementary Materials

A Benchmark In-situ sequencing resource

A.1 Image dataset

We created a benchmark dataset of cells treated with a library of genetic reagents tagged with 186 barcodes of nine digits each. Our plate of cells consists of two rows and three columns of wells. Each well contains about 500000 single cells and 20 million four-channel spots in ISS images of nine cycles. Not all cells have spots, and some cells have more than one spot. Each well is imaged at 316 sites (locations) within the well; we used our in-house pipelines to correct for general microscopic illumination patterns on each image and align the images across cycles in order to correct for small differences in physical plate position on the microscope for each cycle's imaging. The images are subsequently stitched, scaled, and recropped into 100 larger "pseudo-site" (referred as sites in the paper) images. Each pseudo-site's image dimensions are (x:5500, y:5500, channels:4, cycles:9). We make this preprocessed dataset available as a bench-marking resource for developing computational barcode calling methods using ISS images.

A.2 Validation resource for cell-level barcode abundance in a pool of single cells.

As there is no direct ground truth for the barcodes assigned to each image location, we evaluate the barcode calling performance in an indirect way. By applying Next-Generation-Sequencing (NGS) on the pooled screens we can quantify the expressed integrated barcodes. We applied NGS to a separate sample of the same cell population that was placed into the 6 wells of our plate, enabling us to count the number of cells perturbed by each barcode. Abundance of transcripts for these genomically integrated barcodes were captured by kallisto tool. Abundance of barcodes based on any decoding strategy applied on this dataset can be calculated as a post-processing step. We can then compare barcode calling strategies to the NGS abundance measures as the experiments' "perturbation abundance ground-truth". The NGS data is at the bulk level, i.e. a pool of cells are all sequenced together. Because each barcode integrates into the cell's DNA once and only once, and most cells receive only a single integration due to our experimental setup, the NGS information approximates the abundance of cells with a specific transcript or barcode. By contrast, the image-based barcode calling methods read out mRNA spots which can be present in variable copy numbers per cell rather than genomic DNA (which is present in only one copy per cell). For this reason, we cannot expect the number of NGS reads of barcoded cells to linearly correlate with the number of image-based reads of barcoded mRNA transcripts; to assess correlation of our results with NGS data we therefore first need to assign barcode spots to cells to produce cell-level barcode assignments. As explained in Section 4.5 of the paper, for the methods which provide a confidence metric on detected barcodes, we assign each cell with the most confident barcode within that cell. For the methods with no confidence scores on the detected barcodes, we assign the barcode with the largest number of occurrence to each cell. And in the case of multiple barcodes with equal occurrence rate, we simply skip the cell assignment.

A.3 Evaluation Metrics.

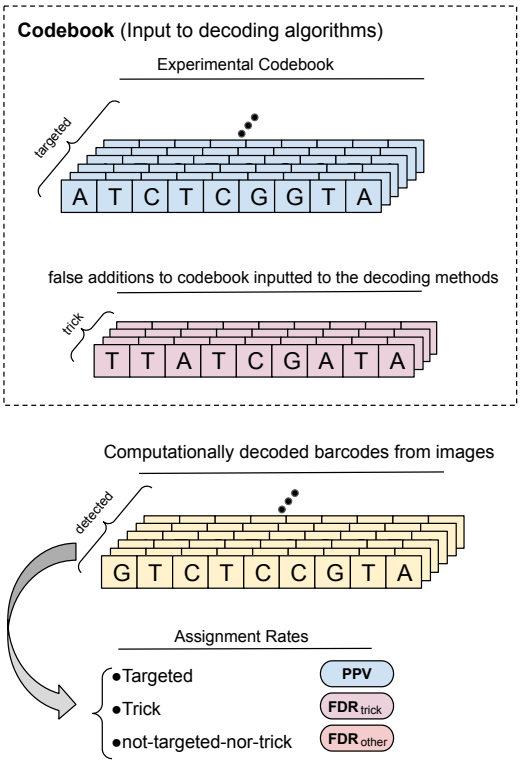
We aim to achieve the highest possible number of cells with a correct barcode assignment. Therefore, the main evaluation metrics are rate of cell recovery and the matches between abundance of cell assignments and the NGS-based barcode abundance:

- **Cell Recovery Rate**

  Recovery rate is defined as the ratio of cell assignments with a targeted barcode over the total number of detected cells by CellProfiler. Note that there are a number of cells that dont recieve any barcode assignments and therefore this number is different than the PPV at the cell level.

- **NGS match**

  As described in Section A.2, NGS-based relative abundance of each barcode in the experiment serves as an indirect ground truth to assess the quality of the relative abundance of the detected barcodes assigned to the cells which also exist in the experimental codebook. The similarity between the abundance of the detected cell-level barcode assignments and the NGS-based abundance of codebook barcodes is measured by R^2 between the two abundance distributions.



Supplementary Figure 2: Schematic illustration of terminologies used for evaluation. Experimental Codebook consists of the reference library of the mRNA barcodes encoded into the cells. False additional *trick* barcodes are added to the experimental codebook as the input to the barcode calling algorithms. Detected barcodes are the set of barcodes derived by computational analysis of the ISS images.

- **False Discovery Analysis**

  The next set of metrics are for false discovery analysis and are calculated at both spot-level and cell-level barcode assignments. Supplementary Figure 2 illustrates the distinction among the various possible types of barcode assignments. The codebook inputted to the decoding algorithms contains two sets of barcodes:

  - **Targeted** barcodes which form the experimental codebook or the experiments reference library of barcodes.
  - **Trick** barcodes are a collection of intentionally fabricated barcodes that, although not part of the original experimental codebook, are added to the codebook inputted to the decoding algorithms. These barcodes, once introduced into the decoding algorithms, serve as an insightful tool for false discovery analysis, aiding in the identification and assessment of potential overfitting issues inherent in the decoding process.

  A decoding method can generally assign spots to any sequence of base letters and therefore there exists a third category of "not-targeted-nor-trick" which include the rest of assignmnets that are not targeted nor trick barcodes.

  - **FDR**: Incorrect assignment rates for two categories of "trick" and "not-targeted-nor-trick" calls are reported as False Discovery Rates (FDR) and are denoted as FDR_{trick} and FDR_{other} respectively.
  - **PPV**: Correct assignment rate refers to the rate of the targeted assignments which is $1 - FDR$. We report this metric as the Positive Predictive Value (PPV) at each spot and cell-level barcode assignments as well.

B PLePI algorithm

The summary of the proposed PLePI algorithm as described in Section 3 and represented in schematic Figure 1.

Algorithm 1 PLePI algorithm

Input: $\mathcal{E}_T, \mathcal{E}_{PI}$
Parameter: τ_m, τ_c
Output: PI enhanced pseudo-labels

Given Teacher's soft decisions $\mathcal{P}_T$ for each sample o in a mini-batch of unlabeled samples, $\mathcal{B}_u \subset \mathcal{D}_u$
while $|\mathcal{B}_u| \neq 0$ **do**
  Sort out samples to three categories.
  if $\max_{k \in K} \mathcal{P}_T(y_o = k) > \tau_c$ **then**
    $\mathcal{O}_C = \mathcal{O}_C \cup \{o\}$
  else if $\max_{k \in K} \mathcal{P}_T(y_o = k) < \tau_m$ **then**
    remove sample form loss calculations
  else
    $\mathcal{O}_M = \mathcal{O}_M \cup \{o\}$
  end if
  for $o_m \in \mathcal{O}_M$ **do**
    perform **fusion** of $\mathcal{E}_{PI}$ into $\mathcal{E}_T$ and
    estimate label k^* for o_m by equation (2)
    for $k \in K$ (the set of top n candidates according to $\mathcal{P}_\mathcal{T}$)
  end for
end while
return Teacher's enhanced pseudo-labeled set for $\mathcal{B}_u$

C Extracting evidence from CLIP

Contrastive Language-Image Pre-training (CLIP) [Radford *et al.*, 2021] is a multi-modal vision and language model trained on 400 million (image, text) pairs to learn visual-linguistic representations. Among various downstream tasks, through zero-shot transfer, the pretrained model is able measure the relative similarity between any given pairs of image and text which were unseen in training phase. We exploit the CLIP model's measured similarities in the form of probabilities, which basically measure the similarity between the model's representations of a given image and text, as the source of privileged information to be fused to the teacher's model decisions in the following way:

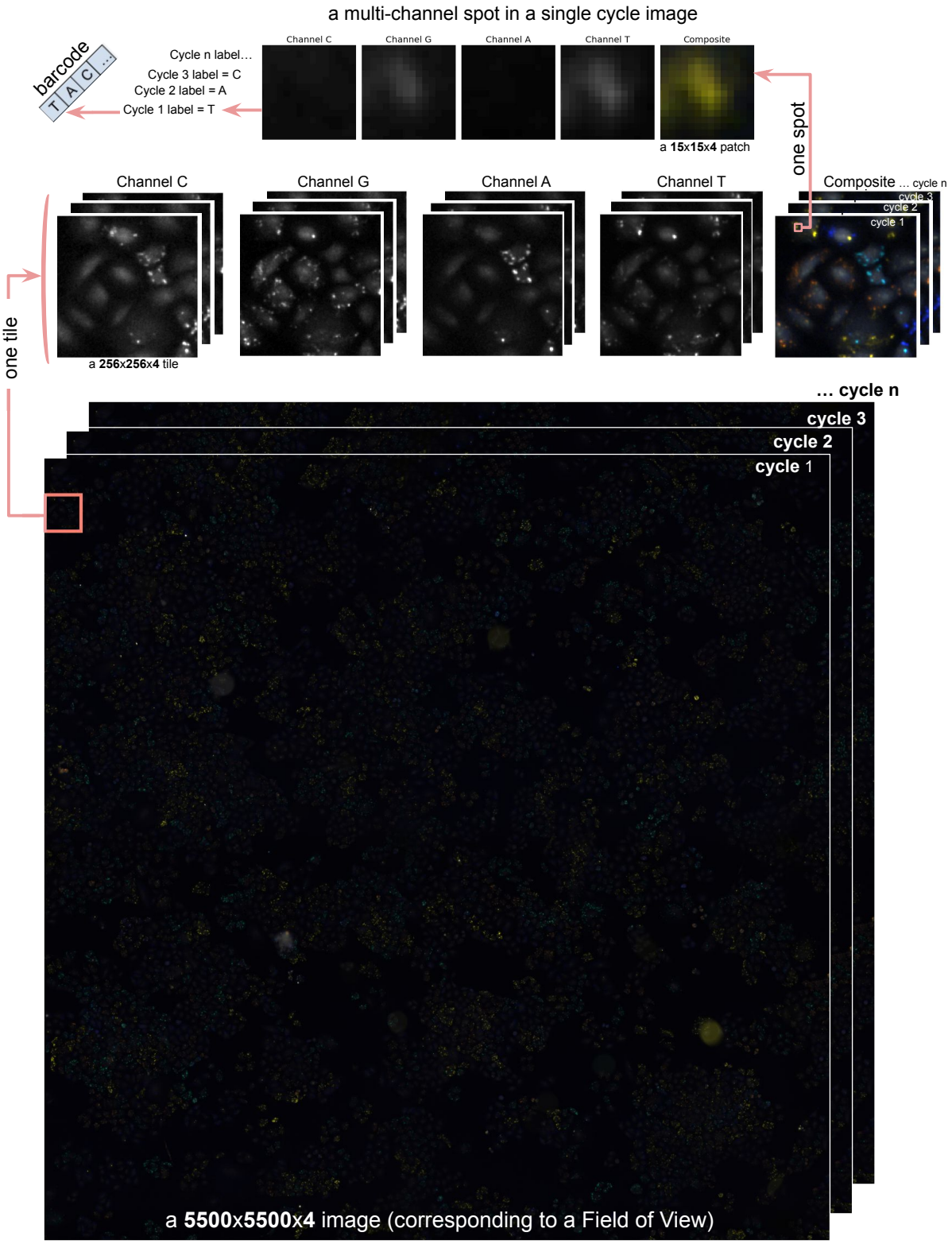
- According to the teacher's model soft predictions ($\mathcal{P}_T$), objects are categorized to three confident $\mathcal{O}_C$, mediocre $\mathcal{O}_M$ and unconfident groups.

- For each $o_m \in \mathcal{O}_M$, privileged information is extracted for the evidence for the set of confident group together with o_m where it belong to category $k \in K$, K being top n candidate categories according to the teacher model $\mathcal{P}_{PI}(o_m \cup \mathcal{O}_C | o_m \in k)$.
- Extracting the above equasion using CLIP as $\mathcal{E}_{PI}$ would result:
  $\mathcal{P}_{CLIP}(text(o_m \cup \mathcal{O}_C), image) | o_m \in k$
- In which $text(o_m \in k \cup \mathcal{O}_C)$ is a sentence formed by concatenation of the class names for each label in the confident and candidate label k for unconfident sample o_m.
- The pseudo-label is then calculated on the fused Teacher and CLIP evidence:

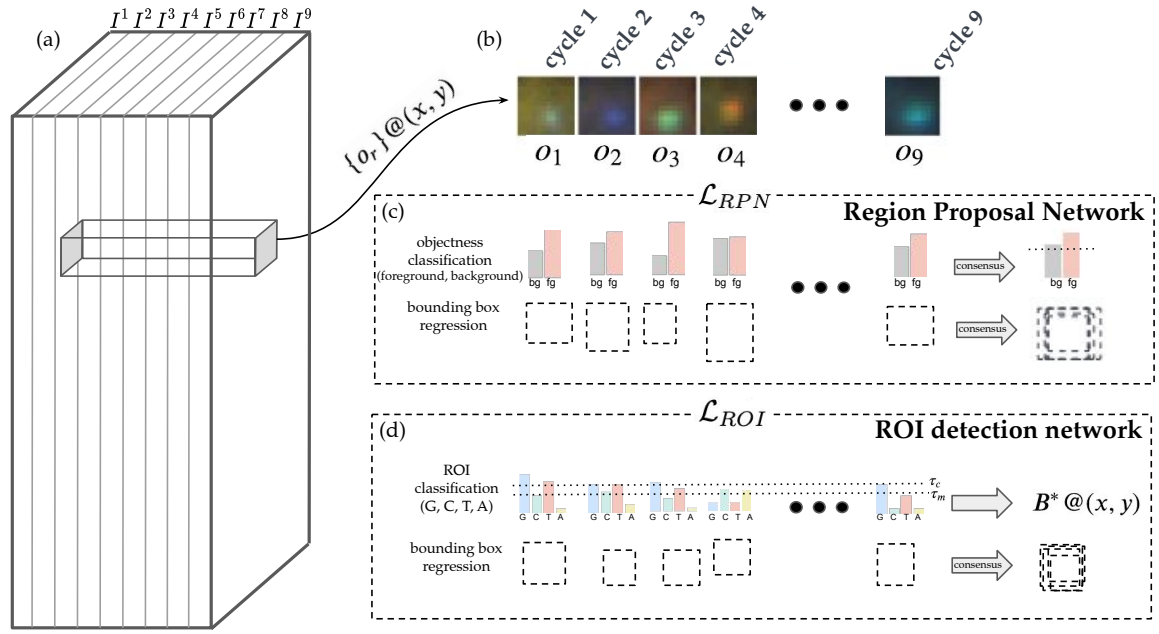
$$k^* = \arg\max_{k \in K} \mathcal{P}_T(o_m \in k) \mathcal{P}_{CLIP}(text(\mathcal{O}_C \cup o_m \in k), image).$$

References

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.



Supplementary Figure 1: Example of real ISS images, showing the 4 color channels (representing C, G, A, T) and multiple cycles that encode a barcode.



Supplementary Figure 3: **PLePI-ISS schematic for a two-stage object detector.** (a) ISS images in a mini-batch of samples. This mini-batch contains nine images $\{I_1, I_2, .., I_9\}$ corresponding to nine rounds of images in a tile of experiment ($N_r = 9$). (b) Nine spots (objects) are extracted from (x, y) location for all images in the mini-batch $\{o_1, o_2, .., o_9\}$. (c) For the specific anchor shown in subfigure a. region proposal network (RPN) generates objectness scores corresponding to the binary classification of foreground versus background. RPN generates bounding box deltas for the proposal anchor corresponding to the bounding box regression task for object localization. The privileged information rooted from organization of samples in a mini-batch dictates the consistency of the objectness evidence across objects at (x, y) location. Therefore, based on consensus (median) teacher soft decisions across the group of objects – as schematically shown in figure – we pseudo-label the objects within a group to be either all in foreground or all in background class and provide those pseudo-labels to the student network. The pseudo-boxes for each object are also calculated by the median of estimated bounding box deltas for detected-as-foreground group of objects and fed to the bounding box regression loss of the student network. (d) A region of interest (ROI) detection network is simultaneously trained for object classification and bounding box regression of a selected set of ROIs. For object classification, objects are divided to two sets of confident and mediocre objects based on the τ_c and τ_m thresholds on the teacher's soft decisions on the class labels. The privileged information – which is the codebook – updates the pseudo-labels of the mediocre objects so that the final barcode is the most probable barcode that also exists in the codebook.