Towards Early Prediction of Human iPSC Reprogramming Success

Abhineet Singh https://orcid.org/0000-0002-5624-8377   asingh1@ualberta.ca
Department of Computing Science, University of Alberta

Ila Jasra https://orcid.org/0000-0002-2842-8565   jasra@ualberta.ca
Alberta Diabetes Institute, University of Alberta

Omar Mouhammed https://orcid.org/0000-0003-3537-9420   mouhamme@ualberta.ca
Alberta Diabetes Institute, University of Alberta

Nidheesh Dadheech https://orcid.org/0000-0002-7155-0486   dadheech@ualberta.ca
Alberta Diabetes Institute, University of Alberta

Nilanjan Ray https://orcid.org/0000-0002-7588-5400   nray1@ualberta.ca
Department of Computing Science, University of Alberta

James Shapiro https://orcid.org/0000-0002-6215-0990   jshapiro@ualberta.ca
Alberta Diabetes Institute, University of Alberta

Abstract

This paper presents advancements in automated early-stage prediction of the success of reprogramming human induced pluripotent stem cells (iPSCs) as a potential source for regenerative cell therapies. The minuscule success rate of iPSC-reprogramming of around 0.01% to 0.1% makes it labor-intensive, time-consuming, and exorbitantly expensive to generate a stable iPSC line since that requires culturing of millions of cells and intense biological scrutiny of multiple clones to identify a single optimal clone. The ability to reliably predict which cells are likely to establish as an optimal iPSC line at an early stage of pluripotency would therefore be ground-breaking in rendering this a practical and cost-effective approach to personalized medicine.

Temporal information about changes in cellular appearance over time is crucial for predicting its future growth outcomes. In order to generate this data, we first performed continuous time-lapse imaging of iPSCs in culture using an ultra-high resolution microscope. We then annotated the locations and identities of cells in late-stage images where reliable manual identification is possible. Next, we propagated these labels backwards in time using a semi-automated tracking system to obtain labels for early stages of growth. Finally, we used this data to train deep neural networks to perform automatic cell segmentation and classification.

Our code and data are available at `https://github.com/abhineet123/ipsc_prediction`.

Keywords: iPSC, microscopy imaging, time-lapse imaging, deep learning, transformers, classification, segmentation, tracking, retrospective labeling

1. Introduction

1.1 Motivation

The goal of this work is to apply machine learning to automate the identification of human iPSCs that show promise for clinical cell therapies in regenerative medicine. IPSCs are

generated by reprogramming a patient's own cells back in time to make more malleable cells with differentiation potential for generating any cells or tissues of interest. This technology has shown great potential for transforming regenerative cell therapies, drug and disease modelling, tissue repair and regeneration, and personalized gene-corrected products. However, the pipeline for iPSC generation, characterization and cell banking is a highly labor-intensive, time-consuming and costly one. The monetary cost of research-grade iPSC line generation is estimated at USD 10,000-25,000 while that of clinical-grade iPSC line is approximately USD 800,000 based on published reports (Huang et al., 2019). The entire process of optimal iPSC line generation and selection can take up to 35 days and requires a further 3 months to produce large scale iPSCs for therapeutic application in patients.

Additionally, quality control techniques for growing iPSCs to limit inter- or intra-patient iPSC line variability, which is currently assessed manually, remain imperfect in large-scale biomanufacturing. The current solution relies on the judgement of an expert cell biologist, who determines precise iPSC induction, confirms pluripotency based on morphological changes and assesses molecular characterization for multiple clones - all tasks that remain highly effort-intensive and subjectively biased. Manual cell quality control therefore cannot be used to scale up the production of iPSCs and derived products for therapeutic applications. An automated method enabling high-throughput surveillance and validation of cell identity, growth kinetics, and morphological features is desirable throughout the entire manufacturing process. The screening is multifold and needed to not only select optimal cells which have been fully converted to iPSCs during reprogramming stage but also to exclude unstable and pseudo iPSC contaminants during the expansion stage. Automating this process using machine learning would therefore be ground-breaking in improving iPSC bioprocess efficiency and yield, thereby drastically reducing the time and cost involved in the generation of iPSC-based products for therapeutic applications. This paper presents some early but promising steps in this direction.

1.2 Background

1.2.1 iPSC Reprogramming

Takahashi and Yamanaka (2006) demonstrated that mouse embryonic or adult fibroblasts can be reprogrammed into pluripotent stem cells by introducing four genes encoding transcription factors, namely Oct3/4, Sox2, Klf4, and c-MYC (Takahashi et al., 2007; Ye et al., 2013). Generated stem cells showed similar morphological or functional behavior as embryonic pluripotent stem cells and were thus termed iPSCs. Soon thereafter, Takahashi et al. (2007) reported directed conversion of human fibroblasts into pluripotent stem cells, termed as human iPSCs. With the discovery of Yamanaka's human iPSC technology, patient-derived stem cells have huge potential in regenerative medicine (Takahashi and Yamanaka, 2013). Human iPSCs show merit not only in delivering any desired cell types for treating degenerative diseases, tissue repairing, disease modeling, and drug screening (Amabile and Meissner, 2009; Yamanaka, 2009), but they also solve two major problems associated with other pluripotent stem cells such as embryonic stem cells (Ye et al., 2013), namely immune tolrenace after transplantation and ethical concerns. However, there still exist technical and biomedical challenges including the risk of teratoma formation and the uncertainty of efficient nuclear reprogramming completeness due to variability and inconsistencies in the

selection of optimal cells (Ye et al., 2013). There are two major problems to be solved before human iPSCs can be applied as a standardized technique, Firstly, manually monitoring the quality of growing iPSC colonies that is currently practiced does not scale. Secondly, only colonies that satisfy clinical good manufacturing practice (GMP) standards need to be identified for use in downstream applications. Hence, there is an urgent need for automated quality control, thereby also lending it an element of objectivity and standardization.

1.2.2 Machine learning in iPSC Recognition

Though many applications of machine learning for iPSC recognition in images have been presented in the literature (Kusumoto et al., 2018; Waisman et al., 2019; Zhang et al., 2019; Hirose et al., 2021; Coronnello and Francipane, 2021; Kusumoto et al., 2022; Lan et al., 2022), there are none that include both detection and classification or use time-lapse imaging, which is the object of this study. To the best of our knowledge, Zhang et al. (2019) presented the method that comes closest to this work though that too differs in several key respects. It utilizes fluorescence imaging and the commercial closed-source IMARIS software to segment cells and it captures 3D shape information that is the basis for extracting morphological features to train the classifier it uses. Our aim is to make open-source cell segmentation possible without fluorescence and with only the 2D pixel data in standard phase-contrast microscpy images.

1.2.3 Deep Learning in Visual Recognition

In the past decade, deep learning (Alzubaidi et al., 2021) has been applied extensively in computer vision (Chai et al., 2021), especially for recognition tasks like image classification (Byerly et al., 2022), object detection (Liu et al., 2020), instance segmentation (Gu et al., 2022), semantic segmentation (Mo et al., 2022), and object tracking (Marvasti-Zadeh et al., 2021; Jiao et al., 2021; Pal et al., 2021). It has likewise seen broad application in medical image analysis (Suganyadevi et al., 2021; Cai et al., 2020) including segmentation in general (Liu et al., 2021a) and cell segmentation in particular (Wen et al., 2022). The latter is the task most relevant to this work, though cell tracking (Ben-Haim and Riklin-Raviv, 2022; Chen et al., 2021; Wang et al., 2020; Lugagne et al., 2020; Ulman et al., 2017) is also important here. More recently, the advent of transformers (Vaswani et al., 2017) has led to significant performance improvements (Liu et al., 2021b) over the convolutional neural network (CNN) based architectures that had been prominent earlier. Liu et al. (2021c) proposed the Swin transformer to improve the original vision transformer (Dosovitskiy et al., 2021) further using shifted windows. This currently appears to be the backbone of choice in most state of the art models, though that might change with the recent introduction of ConvNext (Liu et al., 2022) as a competitive CNN-based alternative. For this project, we needed instance segmentation models, preferably ones that could benefit from the temporal information in time-lapse images. Therefore, we selected top-performing static and video segmentation models (sec. 2.2) with publicly available code from a popular leaderboard (Xiang, 2023). We also searched through the leaderboards on a couple of cell tracking and segmentation benchmark challenges (Ulman et al., 2017; Anjum and Gurari, 2020) but failed to find any models with publicly available code.

2. Methodology

2.1 Data Collection

2.1.1 Cell culturing

Cells were cultured in a Class-II biocontainment compliant lab with manipulation of cells in a sterile environment using a laminar flow hood with high efficiency particulate air filtration. Cells were maintained at 37°C with 5% CO_2 within humidified incubators. Human iPSCs reprogrammed from patient-derived peripheral blood mononuclear cells (PBMCs) were previously established and characterized for pluripotency in the laboratory and cryostored in ultra-low temperature freezers for long-term storage. Before starting the time-lapse imaging, cells from a frozen vial were thawed in a 37°C waterbath and cultured on a matrix-coated 6-well plate in iPSC growth media at a seeding density of 100,000 cells per well according to previously published protocol (Park et al., 2008). Once the cells got attached, fresh media was replenished in each well every day prior to the time-lapse imaging which was initiated on day 6 right after cell-seeding.

2.1.2 Time-lapse imaging

Time-lapse images were captured at 15-minute intervals on a Nikon BioStudio-T microscope using the NIS-Elements cell observation and image analysis software. The images were captured with a 4x lens using a full plate scanning module. A total of 275 images were captured spanning 68.75 hours. The raw images were 10992×10733 pixels or 714 megapixels in size though the usable circular cell culture region in the center comprised only about 85 megapixels with a diameter of 5200 pixels. A sample image is shown in the supplementary. Manual examination showed that images before frame 146 (or 36.5 hours) were unsuitable for our experiments since they contained too many clones, with most being too small and indistinct for reliable labeling. Three of the remaining frames - 155, 186 and 189 - were blurry due to camera shake and had to be discarded too. As a result, a total of 127 frames were used for all experiments.

2.1.3 Annotation

The original 714 megapixels images are too large to be processed directly so they were first divided into several regions of interest (ROIs) (shown in the supplementary) varying in size from 1700×900 to 4333×3833. These were then annotated in 3 stages.

Selective Uncategorized Semi-Automated Labeling The 127 frames were first divided into three sets representing different levels of cell development – 146-200, 201-250 and 251-275. Set-specific ROI sequences were then created, that is, each ROI spanned only one of the three sets instead of all 127 frames. This strategy was chosen to include a good representation of cellular appearance from all stages of development in the labeled data with minimum amount of overall labeling. There were 3, 8 and 6 ROIs from the three sets respectively and these 17 ROIs had a total of 656 frames and 4768 cells.

These were then labeled to mark the locations and pixel-wise masks of cells through a custom-designed graphical labeling tool. This tool integrates the SiamMask tracker (Wang et al., 2018) to semi-automate the labeling process by propagating manually-created masks

into future frames by joint unsupervised segmentation and tracking. Tracking was stopped manually when it started to fail and then restarted from the last frame where it worked, after making any required fixes in the intermediate frames. The labeling tool also supports using a previously trained segmentation model, if available, to automatically generate initial cell candidates that can then be manually modified instead of having to be drawn from scratch, although this capability was not used at this stage. Note that this relatively labor-intensive stage did not require involvement from iPSC detection experts since the labeled cells were not categorized into good and bad.

Exhaustive Automated Labeling A Swin transformer instance segmentation model (Liu et al., 2021c) was first trained on the annotations from the previous stage. Next, ROI sequences spanning all 127 frames were created. These were designed to cover as much of the circular well area containing cells as possible while minimizing overlap between different ROIs. A total of 31 sequences were created by extending 11 of the 17 ROIs from the previous stage to the remaining frames and creating 20 new ones. These sequences had 3937 frames with 22,598 cells in all. More details about this dataset as well as the one from the previous stage are available in the supplementary. Finally, the trained Swin transformer model was used to automatically detect and segment cells in each of these frames.

Categorized Retrospective Labeling An iPSC detection expert first manually categorized the cells in frame 275 from each of the 31 ROIs into good and bad. The two categories were respectively named iPSC and differentiating cell (**DfC**) to reflect their likely future growth outcomes. A semi-automated interactive multi-object tracking tool was then used to propagate these labels backwards in time by tracking each cell line from frame 275 to 146. This process accounted for cell division and fusion events [1] by giving each child cell the same label as the parent in case of division and requiring that all merging cells have the same label in case of fusion. A violation of the latter requirement would have meant that the human expert's labels were incorrect but this never happened which is a sign of their reliability. Note that cell lines that disappear before reaching frame 275 cannot be categorized in this way so these have been excluded from all experiments.

The tracking algorithm was kept simple due to time and computational constraints. Cells were associated between neighbouring frames on the basis of location and shape, similar to the IOU tracker (Bochinski et al., 2017) while likely fusion and division events were detected using heuristics based on the extent of change in the size, shape and location of associated cells, in addition to association failures. The detailed algorithm is included in the supplementary. Due to this simplicity, the retrospective labelling process is currently more time- and labor-intensive than ideal and it took between 10 and 30 minutes to label each ROI sequence depending on density of cells and frequency of division and fusion events.

We have made the code for all three stages publicly available (Singh, 2023) along with the annotated data and trained models to facilitate reproducibility of our results and further work in this domain.

---

1. Note that a division event when going backwards in time is actually a fusion event and vice versa

2.2 Models

We selected two state-of-the-art image classification models to allow head-to-head comparison with the XGBoost classifier-based approach by Zhang et al. (2019) (**XGB**) which is the only existing method in literature that is relevant to this work. The models we chose are based on Swin transformer (Liu et al., 2021c) (**SWC**) and ConvNext (Liu et al., 2022) (**CNC**) architectures. We also selected five instance segmentation models since both cell detection and classification are needed in the absence of fluorescence that was used by Zhang et al. (2019) to identify cells. Two of these are static detectors that process each frame independently and discard any video information. They are both variants of Cascade Mask RCNN (Cai and Vasconcelos, 2021) that differ in their backbone architectures, these being the same as the classifiers chosen above - Swin transformer (Liu et al., 2021c) (**SWD**) and ConvNext (Liu et al., 2022) (**CND**). The remaining three are video detectors that combine information from multiple video frames to make their decisions. One of these - **IDOL** (Wu et al., 2022) - is an online model that only uses information from past frames while the other two - SeqFormer (Wu et al., 2021) (**SEQ**) and **VITA** (Heo et al., 2022) - are batch models that use information from the entire video sequence including both past and future frames. All three video detectors use versions of the Swin transformer backbone. We also experimented with two ResNet (He et al., 2016) variants of VITA but they did not perform well and so have been excluded here. In addition, we trained a Swin transformer semantic segmentation model and tried several ensemble techniques to combine semantic and instance segmentation results but these did not yield significant performance improvements and are therefore likewise excluded.

2.3 Training

Since retrospective labeling is the most time-consuming part of our pipeline, it is desirable to minimize the need to label backwards to as few frames as possible. Therefore, in addition to comparing modern deep learning-based methods with XGB, we also wanted to evaluate the extent to which model performance on early-stage images depends on the lateness of the frames on which it is trained. We constructed two different training datasets to achieve this - an early-stage set with 38 frames from 163 to 202 and a late-stage set with 73 frames from 203 to 275. Models trained on both datasets were evaluated on the 16 frames from 146 to 162. We also had to adapt the method proposed by Zhang et al. (2019) to work with our images since that work computes several of the features using 3D image information which is not available in our case. It turned out that only 7 of the 11 features used there could be suitably approximated with 2D data so we trained XGB using only these 7 features whose details are in the supplementary. All models were trained to convergence using mostly default settings recommended for each model by the respective authors except for minor tinkering with batch sizes, data augmentation strategies and scaling factors to make the models fit in the limited GPU memory available. The classification datasets were constructed from image patches corresponding to the bounding box around each labeled cell with an additional 5 pixel border for context. Models were trained on several different GPU configurations - video detectors: $2 \times$ Tesla A100 40 GB, static detectors: $2 \times$ RTX 3090 24 GB, classifiers: $3 \times$ RTX 3060 12 GB and $3 \times$ GTX 1080 Ti 11 GB.

3. Evaluation

3.1 Classification Metrics

We used standard receiver-operating-characteristic (ROC) curves and the corresponding area-under-curve (AUC) metric to compare the models. Note that head-to-head comparison between detectors and classifiers is difficult since the former both detect and classify cells while the latter only do classification on all the cells in the ground truth (GT). Two types of detector failures need to be accounted for in order to render such a comparison meaningful:

- False Positives (FP): Detections without matching GT cells
    – Misclassification (**FP-CLS**): a GT DfC is detected but misclassified as iPSC
    – Duplicates (**FP-DUP**): the same GT iPSC is detected multiple times and classified each time as iPSC
    – Non-Existent (**FP-NEX**): an iPSC is detected where no GT cell exists (neither iPSC nor DfC)
- False Negatives (FN): GT cells without matching detections
    – Misclassification (**FN-CLS**): a GT iPSC is detected but misclassified as DfC
    – Missing detection (**FN-DET**): a GT iPSC is not detected at all (neither as iPSC nor as DfC)

FP-NEXs were ignored when computing the classification metrics since manual examination[2] showed that virtually all of these corresponded to one of two cases, neither of which is important in our application:

- **FP-NEX-WHOLE**: unlabeled cells whose labels could not be inferred by retrospective labeling (sec. 2.1.3)
- **FP-NEX-PART**: parts of labeled cells mostly involving ambiguously-shaped cells that could plausibly be interpreted as undergoing division or fusion but were labeled as whole cells.

FP-DUPs were also ignored since multiple detections of an iPSC have little impact in our application. Finally, FN-DETs had to be discarded since ROC curves can only be generated by varying the cut-off used in filtering detections based on their confidence values which are unavailable for undetected cells.

In addition to the AUC of the complete ROC curve, we also used partial AUCs (McClish, 1989) with FP thresholds of 0.1%, 1% and 10%. Even a small number of FPs can be extremely detrimental in our application due to the high cost of culturing non-viable cells so that a model that performs better at these FP rates is preferable to another that is better overall but underperforms here.

3.2 Detection Metrics

The exclusion of FN-DETs while computing the classification metrics can make these biased in favour of detectors with high rates of missing cells but high accuracy for the few cells that they do detect. We accounted for this by incorporating the following detection metrics to evaluate and compare only the detectors:

- Frequency of FN-DETs, FP-DUPs and FP-NEXs

---

2. visualizations for all detection failures are included in the supplementary material

- Standard detection metrics (Huang et al., 2017) of average precision (**AP**) and AUC of the recall-precision curve (**RP-AUC**)

3.3 Temporal Metrics

The ability to detect iPSCs as early as possible is crucial to our application. Therefore, we also evaluated the models in each of the 16 test frames one-by-one to judge how their performance varied over time. A model that performs better in earlier frames would be preferable to one that performs better overall but underperforms in earlier frames.

3.3.1 Subsequctial Inference

Video detectors detect cells spanning all the frames in their input video instead of frame-by-frame which is incompatible with temporal evaluation since even detections in early frames are done using information from all 16 frames in the test set. To resolve this, we used incremental inference where the detections for each frame are generated by running the detector on a subsequence comprising only that frame and all of the preceding ones so that information from future frames is not used. For example, detections for frames 1, 2 and 3 are respectively generated by running the detector only on subsequences comprising frames (1), (1, 2) and (1, 2, 3).

We used another variant of subsequential inference to evaluate the impact of the number of frames on the performance of video detectors so we can judge whether patterns in the way that cell boundaries change over time provide useful information about their eventual outcome to these detectors. Here, we divided the 16-frame sequence into a set of non-overlapping subsequences, each with a fixed size, and ran inference on each subsequence independently. For example, with a subsequence size of 2, our 16-frame sequence is divided into 8 subsequences - (1, 2), (3, 4), ..., (15, 16). Detections for all frames in each subsequence are then generated by running on only the frames in that subsequence. We experimented with subsequence sizes of 1, 2, 4 and 8.

3.4 Results

3.4.1 Classification

ROC curves for both early and late-stage models are shown in the top row of Fig. 1. It can be seen that the early models perform much better than the late ones, which is expected since their training images are much more similar to the test images. However, the extent of this difference does indicate that cell-appearance changes so much over the course of just 9.5 hours (temporal gap between early and late-stage training images) that long-term retrospective labelling right to the very early-stage images is likely to be essential to generate training data for models that can do early-stage iPSC detection reliably. The significant performance advantage of deep learning models over XGB is also apparent here. The main shortcoming of XGB seemed to be its inability to handle class imbalance in the training set. Since about 75% of all cells were DfCs, XGB apparently learnt to classifty nearly all cells as DfCs. We tried all standard techniques to handle class imbalance (Krawczyk, 2016) but this is the best performance we could get from it.
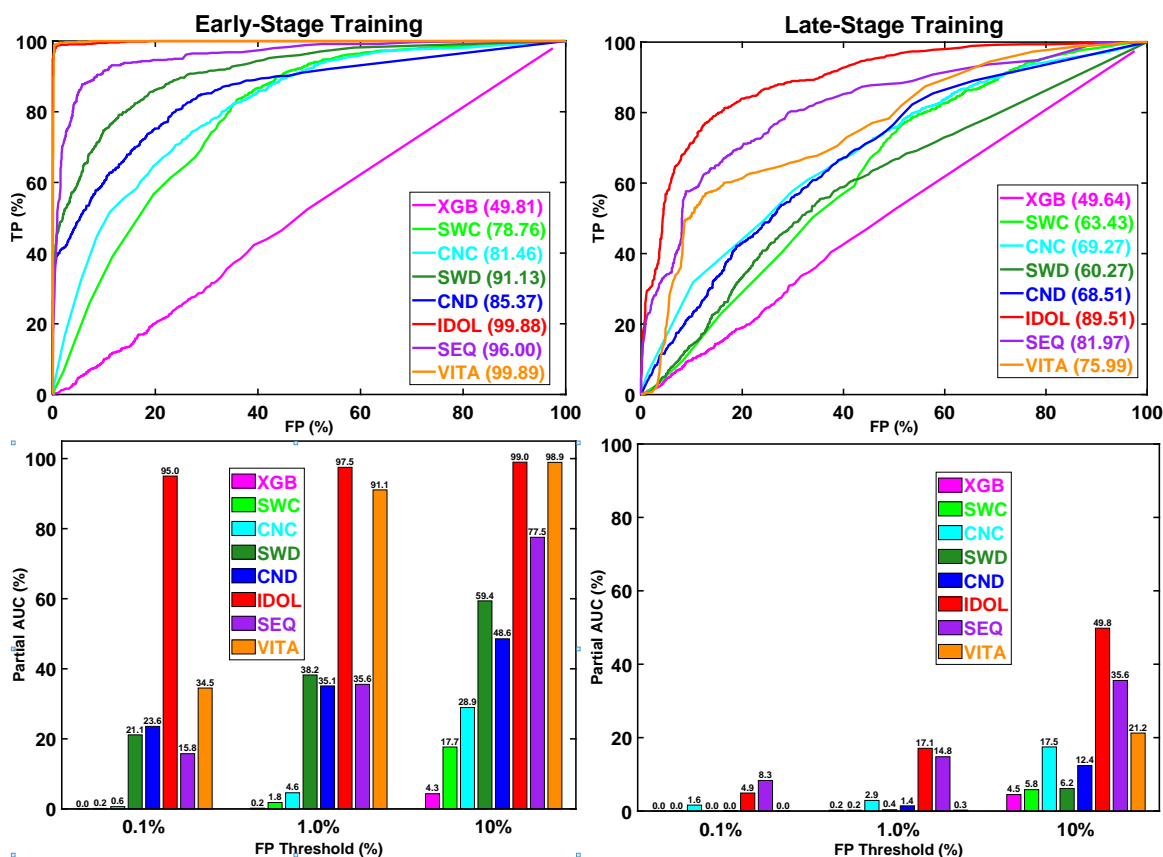
Figure 1: Classification metrics for (left) early and (right) late-stage models. Top: ROC curves (respective AUC values are in the legend); bottom: partial AUCs. Please refer sec. 2.2 for model acronyms.

Further, all the video detectors are consistently better than the static ones which seems to confirm the human experts' supposition that temporal information is crucial for making good predictions. Also, the static detectors do outperform the classifiers but only significantly so in the early-stage case. Since the precise shape of cell boundaries is not available to the classifiers, this lends some weight to the additional supposition that cell-shape is important for recognizing iPSCs. However, as already noted above, the cell-shapes change too rapidly for this information to be generalizable from the later stages to the earlier ones. Finally, IDOL turns out to be the best model overall even though it is the smallest and fastest of the three video detectors while the much larger VITA shows a susceptibility to overfitting in its sharp decline between the two cases. This kind of overfitting is exhibited by both the static detectors too, though it is more strongly marked in case of Swin transformer.

Partial AUCs are shown in the bottom row of Fig. 1. Relative performance between the models is broadly similar to that for overall AUC though IDOL shows greater performance advantage over other models in the early-stage case especially for 0.1% FP. The detectors also show greater improvement over the classifiers in these high-precision scenarios. We
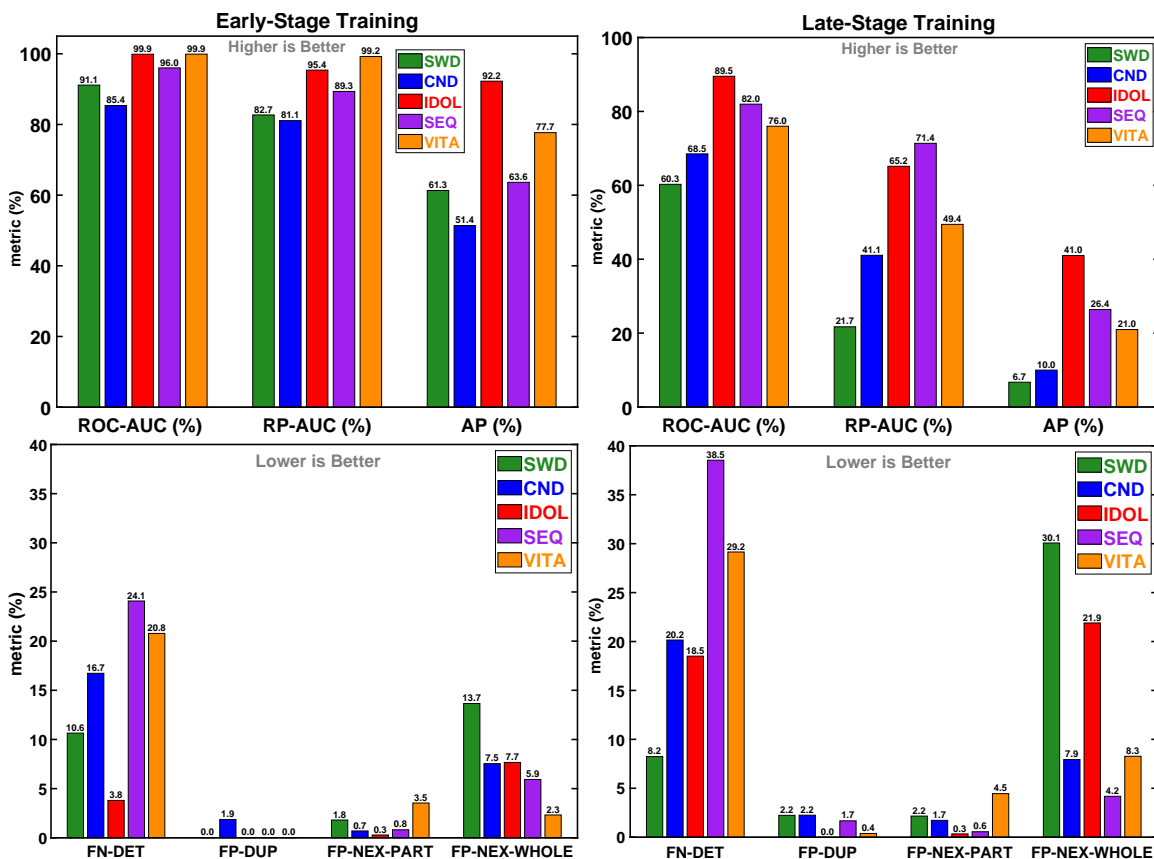
398

Figure 2: Detection metrics for (left) early and (right) late-stage models. ROC-AUC is included for comparison.

also analysed the temporal evolution of partial AUCs by evaluating these frame-by-frame to generate 3D plots with time dimension on the Z-axis (included in the supplementary material) but could not find any useful patterns beyond those apparent in these 2D plots.

3.4.2 DETECTION

As shown in Fig. 2, relative detection performance is mostly consistent with classification accuracy, with IDOL still being the best model overall. IDOL also shows the smallest drop in performance between ROC-AUC and AP while the two static detectors show the largest drops amounting to nearly 10-fold and 7-fold respectively for SWD and CND in the late-stage case. Among the video detectors, VITA has the largest drop, especially in the late-stage case which underlines its overfitting. Further, both SEQ and VITA have very high FN-DET rates that are also at odds with their relatively good classification performance. Conversely, SWD has much lower rates of FN-DETs than its ROC-AUCs would suggest, especially in the late-stage case where it outperforms all other models by a significant margin, though this does come at the cost of a corresponding rise in FP rates. We can also note that this increase in FPs is dominated by only one subtype, namely FP-NEX-
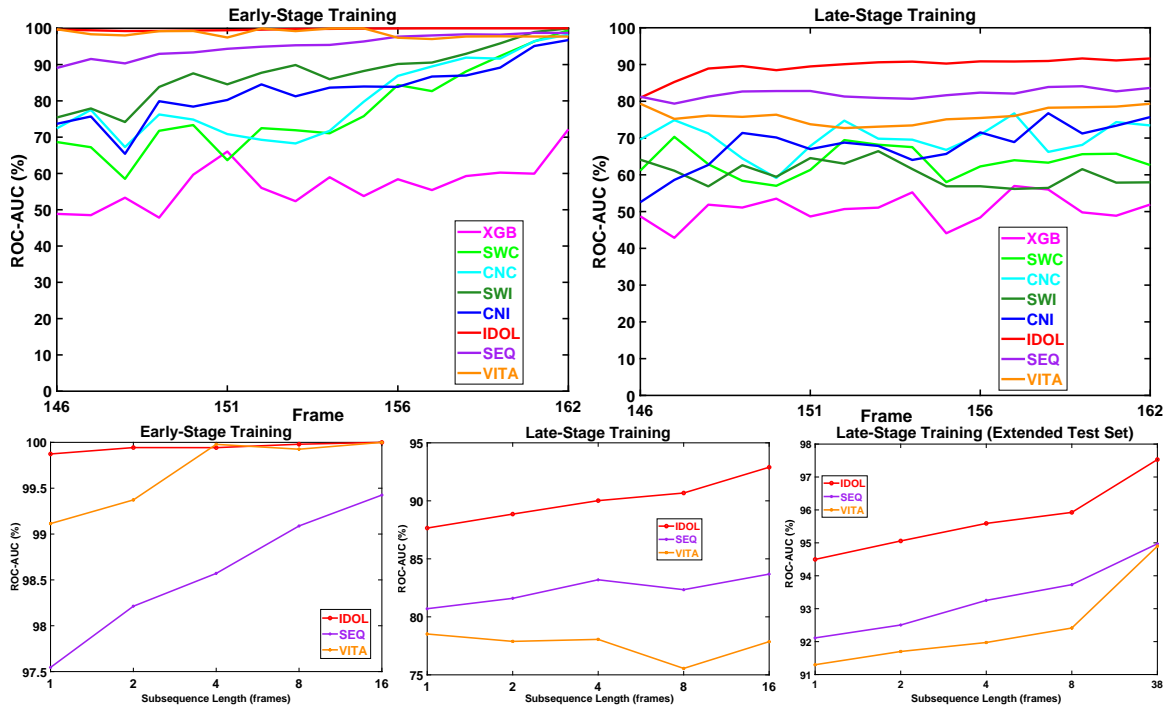
Figure 3: Temporal metrics - top: frame-wise ROC-AUCs for (left) early and (right) late-stage models; bottom: subsequential ROC-AUCs for video detectors - left and center plots show early and late-stage models tested on the standard test set of frames 146 - 162 while the right one shows the latter tested on an extended test set with frames 146 - 201. Note the curtailed and variable Y-axis ranges in the subsequential ROC-AUC plots.

WHOLE, while FP-DUP and FP-NEX-PART show little change, not only for SWD but also for IDOL and VITA. It appears that either due to the greater range of cell-appearances in the late-stage dataset (owing to its greater size) or larger disparity in cell-appearance with respect to the test set, these models learnt to detect a lot more of the unlabeled cells than their early-stage counterparts, which were therefore better at discriminating between the unlabeled and labeled cells.

3.4.3 TEMPORAL

Frame-wise AUCs are shown in the top row of Fig. 3. Somewhat contrary to expectation, AUC does not show consistent increase with time even though test images are becoming more similar to the training images. In fact, many of the early-stage models do show a weak upward trend while late-stage ones do not show any. This is particularly unexpected since the latter perform significantly worse overall and show signs of overfitting, which should lead to a more strongly marked increase in accuracy as resemblance between the test and training images increases. A possible explanation might be that the late-stage training

images are just too far away from the test set for any intra-set variations in test images to make enough of a difference in their resemblance to training images to benefit the models.

Bottom row of Fig. 3 shows the impact of subsequence length on the video detectors. These plots include only ROC-AUC but detection metrics showed similar patterns and have thus been relegated to the supplementary. In order to incorporate the longest-term temporal information possible with our dataset, we also tested the late-stage model on an extended test set comprising frames 146 to 201. Unfortunately, there is a visual discontinuity between frames 184 and 185 due to which this 54-frame sequence had to be divided into two subsequences - one with 38 frames from 146 to 184 and another with 16 frames from 185 to 201 - so that the longest subsequence length was 38 instead of 54. There is indeed a general upward trend with subsequence length but it is much weaker than might be expected. Early-stage models showed negligible impact of subsequence length while the greatest overall gains were 5.2% in the second case and 3.6% in the third case, achieved respectively by IDOL and VITA. Neither of these seem sufficient considering the 4 to 9 hours worth of extra temporal information available to the models. This might indicate that changes in cell-appearance over time are not as useful for recognizing iPSCs as supposed by experts. More likely, it might mean that existing video detectors are simply not good enough to exploit this information sufficiently well and better long-term models are needed.

4. Conclusions

This paper presented a labeling, training and evaluation pipeline along with baseline performance results for early-stage prediction of iPSC reprogramming outcome to help select the best quality clones. These are still early days of research in this domain and it is difficult to say how practical such an automation can be. While it is clear that deep learning models hold significant advantage over previous methods, they also show signs of overfitting and it is unclear how much training data would be needed to overcome such issues.

5. Future Work

We are currently working to improve the tracking algorithm in the retrospective labeling system to make the process faster and less tedious since that is the current bottleneck in the labelling pipeline. We are trying out several state-of-the-art real-time multi object trackers, especially those specialized for tracking cells (Chen et al., 2021; Wang et al., 2020; Lugagne et al., 2020), to replace the simple IOU based algorithm used in this work. We are also looking into ways to improve this algorithm by exploiting recursive parent-child relationships in our labels that can be used to construct hierarchical tree-like structures. Transformer architectures that support such structures (Wang et al., 2019; Ahmed and Mercer, 2021; Wang et al., 2022; Zhong et al., 2022) might help to incorporate cell division and fusion events directly into the algorithm to not only improve the cell association reliability but also reduce the incidence of false positives in the detection of these events, which is the most time-consuming and tedious aspect of this process. A recent method uses graph neural networks to exploit these structures (Ben-Haim and Riklin-Raviv, 2022) for cell tracking. We are trying to incorporate it into our pipeline and also improve it further using transformers. If

these efforts prove successful, we hope to make long term retrospective labelling spanning multiple weeks feasible so any such demands for data can be met.

We are also exploring ways to better exploit long-term patterns in the evolution of cell-appearance over time to improve prediction quality since the current ability of video detectors in this regard does not appear to commensurate with the importance that human experts attach to this information. This discrepancy might be due to both the limited temporal span of our data or the models being unable to benefit from this information because they do not learn from sufficiently wide temporal windows. We hope to resolve the former by extending our data from its current span of only 3 days to as much as 5 weeks once the retrospective labeling system have been improved enough to make such long-term labeling feasible. The latter can be addressed by using more memory-efficient models since increasing the size of temporal windows for training is an extremely memory expensive process and our computational resources are insufficient to train existing models with wider windows.

Another related issue we are trying to address is that of the black-box nature of deep learning which makes it very difficult for the medical experts in our team to figure out the reasons behind particular failures of the models. It makes it equally difficult for us to incorporate their suggestions to fix such failures. We are hoping that recent advances in the field of interpretable deep learning (Li et al., 2021; Samek et al., 2021; Molnar, 2020; Samek et al., 2020, 2019), especially with regard to transformers (Ali et al., 2022), video processing (Anders et al., 2019) and medical imaging (Brima and Atemkeng, 2022), might help with these limitations.

Acknowledgments

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Mahtab Ahmed and Robert E. Mercer. Encoding dependency information inside tree transformer. *Proceedings of the Canadian Conference on Artificial Intelligence*, 2021.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 435–451, 17–23 Jul 2022.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Q. Al-Dujaili, Ye Duan, Omran Al-Shamma, Jesus Santamaría, Mohammed Abdulraheem Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 2021.

Giovanni Amabile and Alexander Meissner. Induced pluripotent stem cells: current progress and potential for regenerative medicine. *Trends Mol. Med.*, 15(2):59–68, February 2009.

Christopher J Anders, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Understanding patch-based learning of video data by explaining predictions. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 297–309, 2019.

Samreen Anjum and Danna Gurari. CTMC: Cell tracking with mitosis detection dataset challenge. *CVPR Workshops*, pages 4228–4237, 2020.

Tal Ben-Haim and Tammy Riklin-Raviv. Graph Neural Network for Cell Tracking in Microscopy Videos. In *ECCV*, 2022.

Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. *AVSS*, pages 1–6, 2017.

Yusuf Brima and Marcellin Atemkeng. What do deep neural networks learn in medical images? *arXiv preprint arXiv:2208.00953*, 2022.

Adam Byerly, Tatiana Kalganova, and Richard Ott. The current state of the art in deep learning for image classification: A review. In *Sai*, 2022.

Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of Translational Medicine*, 8, 2020.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *TPAMI*, 43(5):1483–1498, 2021.

Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021. ISSN 2666-8270.

Yuqian Chen, Yang Song, C. Zhang, Fan Zhang, Lauren J. O'Donnell, Wojciech Chrzanowski, and Weidong (Tom) Cai. Celltrack R-CNN: A novel end-to-end deep neural network for cell segmentation and tracking in microscopy images. *ISBI*, pages 779–782, 2021.

Claudia Coronnello and Maria Giovanna Francipane. Moving towards induced pluripotent stem cell-based therapies with artificial intelligence and machine learning. *Stem Cell Reviews and Reports*, 18:559 – 569, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Wenchao Gu, Shuang Bai, and Ling-Yun Kong. A review on 2d instance segmentation based on deep neural networks. *Image Vis. Comput.*, 120:104401, 2022.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. 2022.

Takuya Hirose, Jun'ichi Kotoku, Fujio Toki, Emi K. Nishimura, and Daisuke Nanba. Label-free quality control and identification of human keratinocyte stem cells by deep learning-based automated cell tracking. *Stem Cells (Dayton, Ohio)*, 39:1091 – 1100, 2021.

Ching-Ying Huang, Chun-Lin Liu, Chien-Yu Ting, Yueh-Ting Chiu, Yu-Che Cheng, Martin W Nicholson, and Patrick C H Hsieh. Human iPSC banking: barriers and opportunities. *J. Biomed. Sci.*, 26(1):87, October 2019.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara Balan, Alireza Fathi, Ian S. Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin P. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CVPR*, 2017.

Licheng Jiao, Dan Wang, Yidong Bai, Puhua Chen, and Fang Liu. Deep learning in visual tracking: A review. *IEEE transactions on neural networks and learning systems*, 2021.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5:221–232, 2016.

Dai Kusumoto, Mark Lachmann, Takeshi Kunihiro, Shinsuke Yuasa, Yoshikazu Kishino, Mai Kimura, Toshiomi Katsuki, Shogo Itoh, Tomohisa Seki, and Keiichi Fukuda. Automated deep learning-based system to identify endothelial cells derived from induced pluripotent stem cells. *Stem Cell Reports*, 10:1687 – 1695, 2018.

Dai Kusumoto, Shinsuke Yuasa, and Keiichi Fukuda. Induced pluripotent stem cell-based drug screening by use of artificial intelligence. *Pharmaceuticals*, 15, 2022.

Yiqing Lan, Nannan Huang, Yiru Fu, Ke Liu, He Zhang, Yuzhou Li, and Sheng Yang. Morphology-based deep learning approach for predicting osteogenic differentiation. *Frontiers in Bioengineering and Biotechnology*, 9, 2022.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64:3197 – 3234, 2021. URL `https://api.semanticscholar.org/CorpusID:232290756`.

Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *IJCV*, 128(2):261–318, 2020. ISSN 1573-1405.

Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 2021a.

Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE transactions on neural networks and learning systems*, PP, 2021b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, pages 9992–10002, 2021c.

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.

Jean-Baptiste Lugagne, Haonan Lin, and Mary J. Dunlop. DeLTA: Automated cell segmentation, tracking, and lineage reconstruction using deep learning. *PLOS Computational Biology*, 16(4):1–18, 04 2020.

Seyed Mojtaba Marvasti-Zadeh, Li Cheng, Hossein Ghanei-Yakhdan, and Shohreh Kasaei. Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–26, 2021.

Donna K McClish. Analyzing a portion of the roc curve. *Medical Decision Making*, 9:190 – 195, 1989.

Yuji Mo, Y. Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493: 626–646, 2022.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Sankar K. Pal, Anima Pramanik, Jhareswar Maiti, and Pabitra Mitra. Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*, 51:6400 – 6429, 2021.

In-Hyun Park, Paul H Lerou, Rui Zhao, Hongguang Huo, and George Q Daley. Generation of human-induced pluripotent stem cells. *Nat. Protoc.*, 3(7):1180–1186, 2008.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, Klaus-Robert Müller, et al. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*, 2, 2020.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109:247–278, 2021. URL `https://api.semanticscholar.org/CorpusID:232046495`.

Abhineet Singh. iPSC prediction with deep learning. online: `https://github.com/abhineet123/ipsc_prediction`, 2023.

S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11:19 – 38, 2021.

Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, August 2006.

Kazutoshi Takahashi and Shinya Yamanaka. Induced pluripotent stem cells in medicine and biology. *Development*, 140(12):2457–2461, June 2013.

Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872, November 2007.

Vladimir Ulman, Martin Mavska, and Klas E. G. Magnusson et al. An objective comparison of cell tracking algorithms. *Nature methods*, 14:1141 – 1152, 2017.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

Ariel Waisman, Alejandro La Greca, Alan M Möbbs, María Agustina Scarafía, Natalia L Santín Velazque, Gabriel Neiman, Lucía N Moro, Carlos Luzzani, Gustavo E Sevlever, Alejandra S Guberman, and Santiago G Miriuka. Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. *Stem Cell Reports*, 12(4):845–859, April 2019.

Junjie Wang, Xiaohong Su, Lingling Zhao, and Jun Zhang. Deep reinforcement learning for data association in cell tracking. *Frontiers in Bioengineering and Biotechnology*, 8, 2020.

Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2018.

Wenhan Wang, Kechi Zhang, Ge Li, Shangqing Liu, Zhi Jin, and Yang Liu. A tree-structured transformer for program representation learning. *ArXiv*, abs/2208.08643, 2022.

Yau-Shian Wang, Hung yi Lee, and Yun-Nung (Vivian) Chen. Tree transformer: Integrating tree structures into self-attention. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Tingxi Wen, Binbin Tong, Y. Liu, Ting Pan, Yu Du, Yuping Chen, and Shanshan Zhang. Review of research on the instance segmentation of cell images. *Computer methods and programs in biomedicine*, 227, 2022.

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, 2021.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Loddon Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022.

Yu Xiang. PapersWithCode - Instance Segmentation. online: `https://paperswithcode.com/task/instance-segmentation`, 2023.

Shinya Yamanaka. A fresh look at iPS cells. *Cell*, 137(1):13–17, April 2009.

Lei Ye, Cory Swingen, and Jianyi Zhang. Induced pluripotent stem cells and their potential for basic and clinical sciences. *Curr. Cardiol. Rev.*, 9(1):63–72, February 2013.

Haishan Zhang, Ximing Shao, Yin Peng, Yanning Teng, Konda Mani Saravanan, Huiling Zhang, Yanjie Wei, and Hongchang Li. A novel machine learning based approach for ips progenitor cell identification. *PLoS Computational Biology*, 15, 2019.

Shuhan Zhong, Sizhe Song, Guanyao Li, and Shueng Han Gary Chan. A tree-based structure-aware transformer decoder for image-to-markup generation. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.