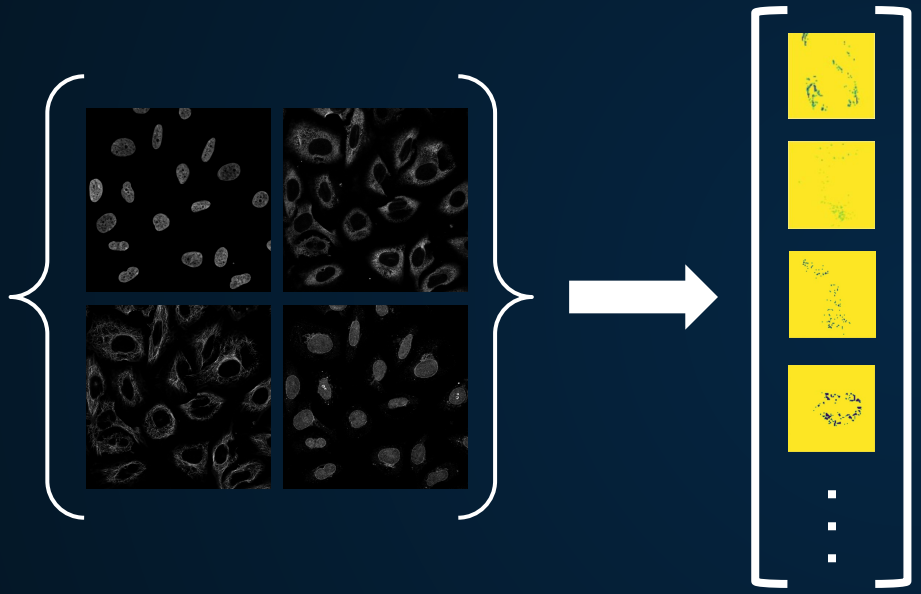
Human Protein Atlas Single Cell Classification

Samuel Golden, Theo Guidroz, Prince Hodonou

TABLE OF CONTENTS



Problem Description

Given images of cells, and *image-level* labels of the proteins being stained for in those cells; segment the image into individual cells and train a model to predict where specific proteins are within new images of cells.

Project Goals

Goal 1

Segmentation for dividing cells

Preprocessing stage for dividing cells

Goal 2

Single Cell Classification

The prediction task is to look at images of the same type and predict the labels of each individual cell within the images

Data

Source of data: Human Protein Atlas Kaggle competition dataset
Description of data: 4 color channels of a group of cells

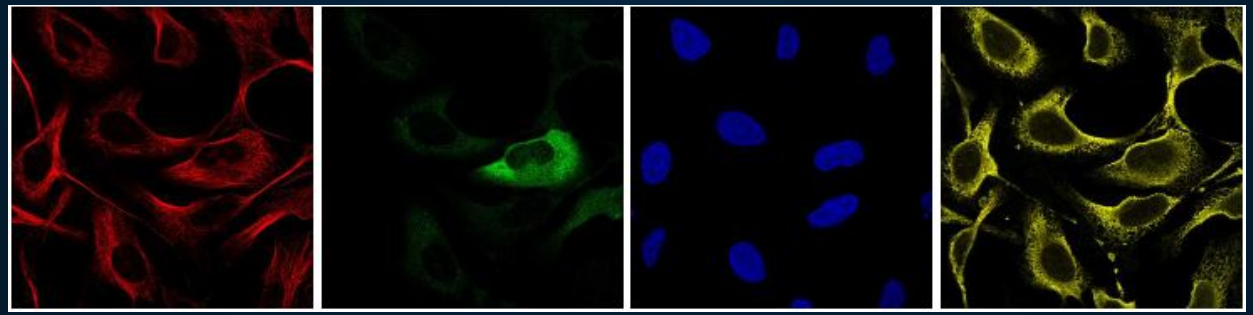
- Red: Microtubules
- Blue: Nucleus
- Yellow: Endoplasmic reticulum

Same for every sample; serve as "markers" for model

- Green: Protein of interest

Different for every sample; basis for model predictions

Label: 19 different classes

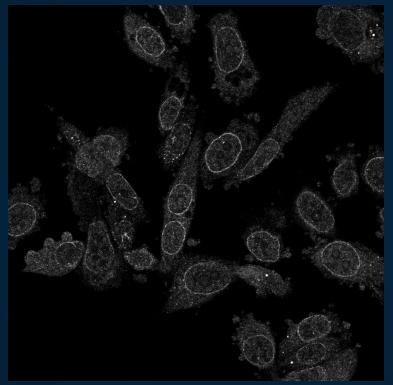


Showing each channel

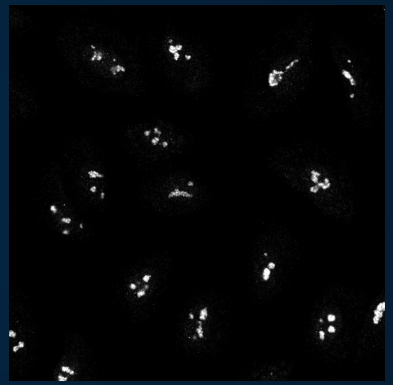
Data

POSSIBLE CLASS LABELS ARE ANY COMBINATION OF:

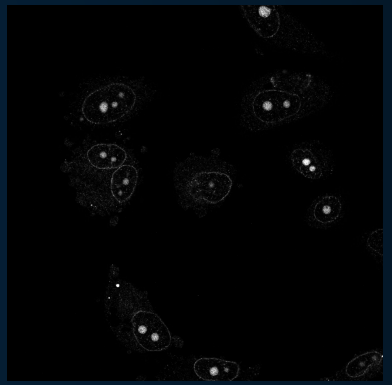
0. Nucleoplasm
1. Nuclear membrane
2. Nucleoli
3. Nucleoli fibrillar center
4. Nuclear speckles
5. Nuclear bodies
6. Endoplasmic reticulum
7. Golgi apparatus
8. Intermediate filaments
9. Actin filaments
10. Microtubules
11. Mitotic spindle
12. Centrosome
13. Plasma membrane
14. Mitochondria
15. Aggresome
16. Cytosol
17. Vesicles and punctate cytosolic patterns
18. Negative



label 1: nuclear membrane



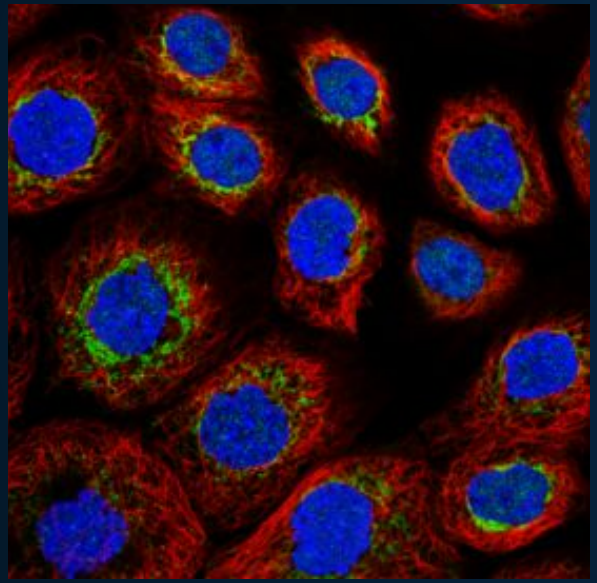
label 2: nucleoli



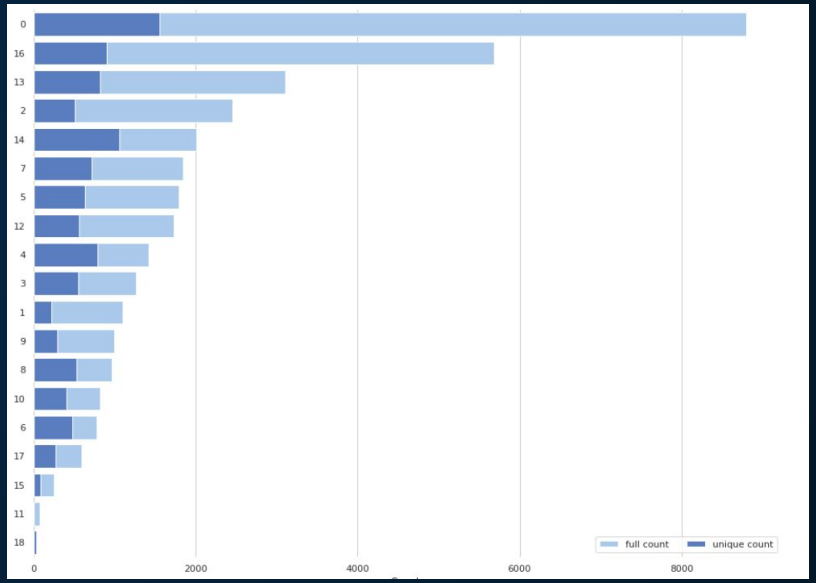
both labels 1 ***and*** 2

"weak labeling"

Data



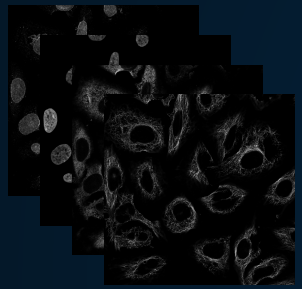
Combining the Red, Green & Blue Channel



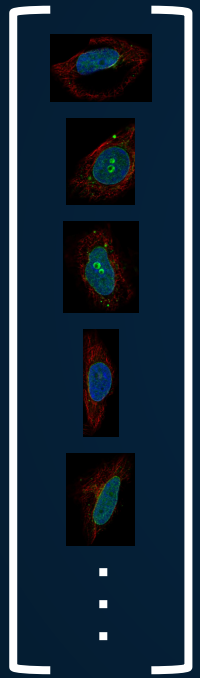
Visualization of unbalanced data

End-to-End Model Architecture

Original 4-channel images of cells

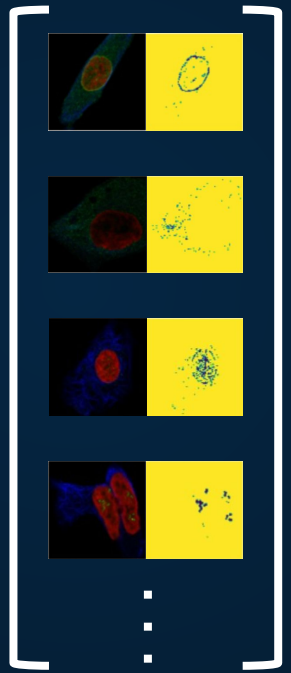


HPA Cell Segmentation Network



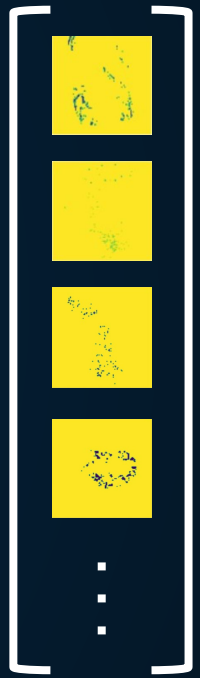
Non-uniform segmented cells

PREPROCESSING



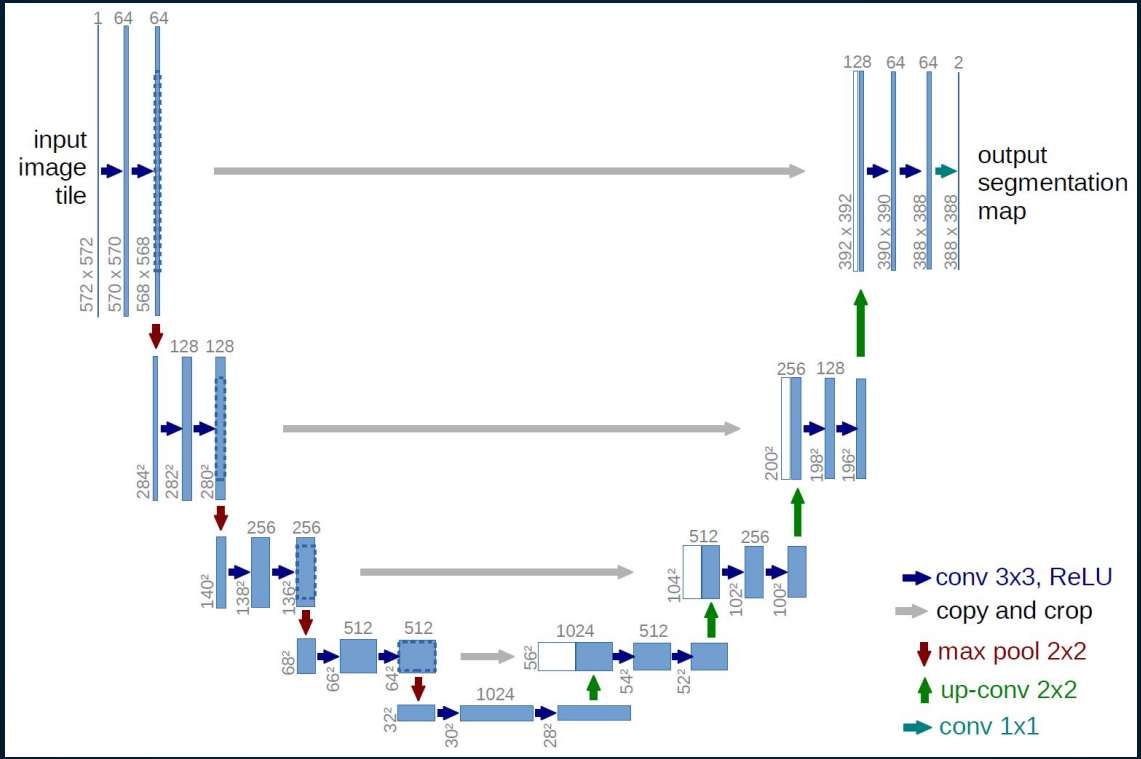
Uniform segmented cells and class mask from green channel (training labels)

U-net

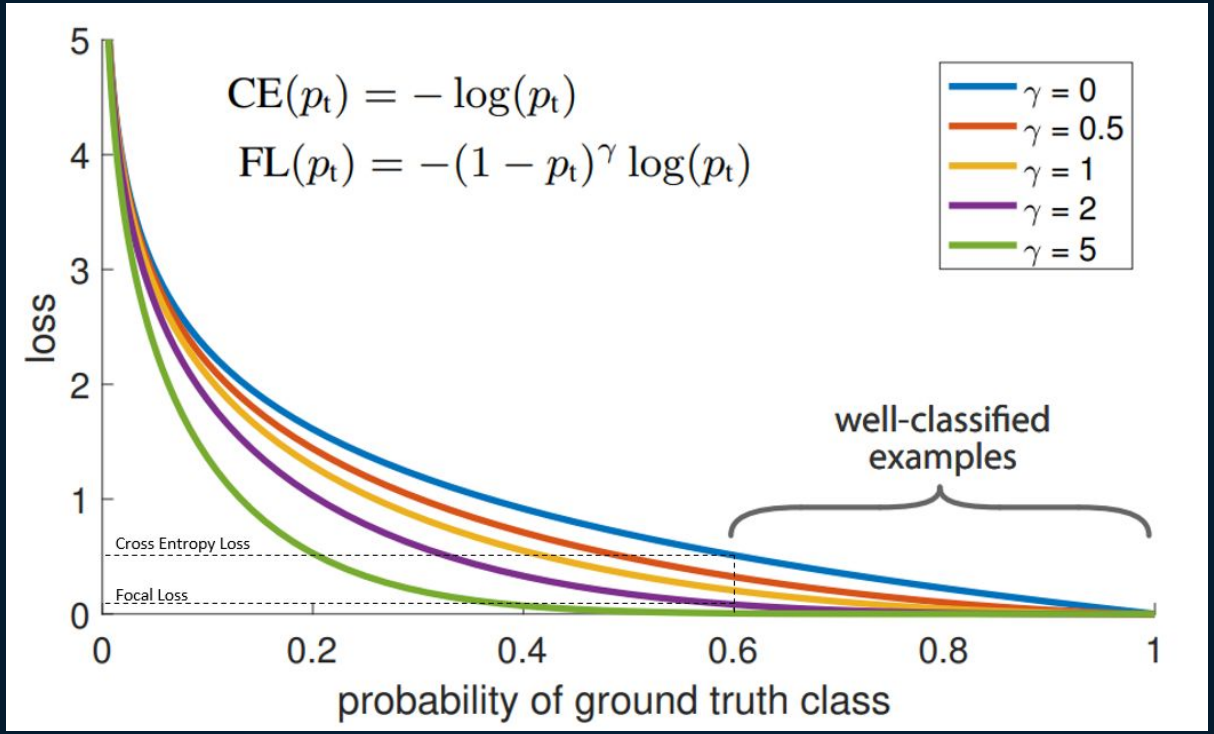


Mask predictions (each 'pixel' is the class prediction)

U-net Architecture



Focal Loss

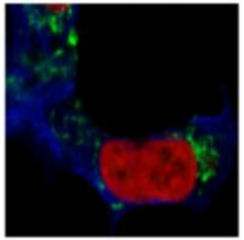

$\mathrm{CE}(p_t) = -\log(p_t)$
$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$
$\gamma = 0$
$\gamma = 0.5$
$\gamma = 1$
$\gamma = 2$
$\gamma = 5$
loss
probability of ground truth class
well-classified examples
Cross Entropy Loss
Focal Loss
0
1
2
3
4
5
0
0.2
0.4
0.6
0.8
1

RESULTS

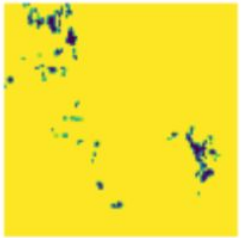
Labels with more data

True label: {14: 363, 18: 16021}
Predicted label: {9: 52, 13: 1, 14: 269, 17: 2, 18: 16060}

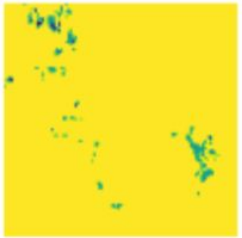
original RGB image



True mask

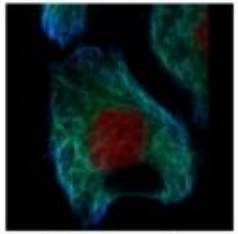


Predicted Mask



True label: {10: 276, 18: 16108}
Predicted label: {10: 260, 18: 16124}

original RGB image



True mask

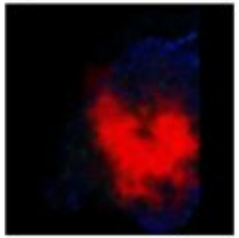
Predicted Mask

RESULTS

Labels with lack of data

True label: {11: 311, 18: 16073}
Predicted label: {0: 20, 10: 27, 12: 40, 16: 40, 18: 16257}

original RGB image

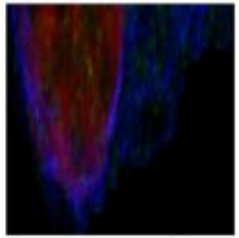


True mask

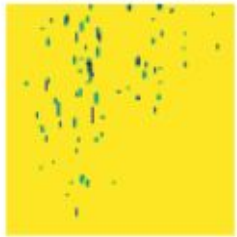
Predicted Mask

True label: {15: 234, 18: 16150}
Predicted label: {1: 115, 8: 73, 10: 7, 18: 16189}

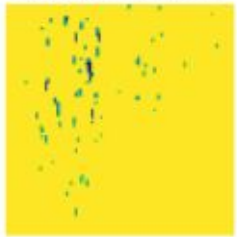
original RGB image



True mask



Predicted Mask



Future Improvement

1. Data augmentation for classes with relatively small amount of data
2. Training the model for multiclass images
3. Work on the submission portion for the Kaggle competition

References

Unet - https://arxiv.org/pdf/1505.04597.pdf

Focal Loss - https://arxiv.org/pdf/1708.02002.pdf

Kaggle Challenge - https://www.kaggle.com/c/hpa-single-cell-image-classification

Kaggle Dataset - https://www.kaggle.com/c/hpa-single-cell-image-classification/data