A TRAINING DETAILS

A.1 DETAILS ON REWARD FUNCTION

We assign a positive reward 1.0 to the agent if an action results in a target hole with a CTF value less than 6.0Å and 0.0 otherwise. The agent also receives a negative reward depending on the operational cost associated with a hole visit. Let $\mathcal{P}_s, \mathcal{Q}_s, \mathcal{G}_s$ be the patch, square and grid index of the hole s. In the end, all the possible rewards and their corresponding conditions are given by

$$r\left(s_i, a_i\right) = \begin{cases} 1.0 & \text{if } \operatorname{ctf}\left(s_{i+1}\right) < 6.0 \ \& \ \mathcal{P}_{s_i} = \mathcal{P}_{s_{i+1}} \\ 0.57 & \text{if } \operatorname{ctf}\left(s_{i+1}\right) < 6.0 \ \& \ \mathcal{P}_{s_i} \neq \mathcal{P}_{s_{i+1}} \ \& \ \mathcal{Q}_{s_i} = \mathcal{Q}_{s_{i+1}} \\ 0.23 & \text{if } \operatorname{ctf}\left(s_{i+1}\right) < 6.0 \ \& \ \mathcal{Q}_{s_i} \neq \mathcal{Q}_{s_{i+1}} \ \& \ \mathcal{G}_{s_i} = \mathcal{G}_{s_{i+1}} \\ 0.09 & \text{if } \operatorname{ctf}\left(s_{i+1}\right) < 6.0 \ \& \ \mathcal{G}_{s_i} \neq \mathcal{G}_{s_{i+1}} \\ 0.0 & \text{otherwise} \end{cases}$$

where $s_{i+1} = T(s_i, a_i)$.

A.2 HYPERPARAMETERS

There are three hyperparameters for the training of *FixMatch* . **uratio** controls the ratio between the number of samples from labeled data and the number of samples from unlabeled data in each batch. **ulb_loss_ratio** is the coefficient of the unsupervised loss. The two hyperparameters are set to 4 and 5.0 respectively. **p_cutoff** (τ in 1) controls minimum confidence it requires to be considered for the unsupervised loss, which is set to 0.8.

The initial classifier is a Resnet 18, trained under learning rate 0.01 with a cosine learning rate scheduler, dropout rate 0.5, batch size 64 for 200 episodes. The in-loop fine-tuning is trained with a learning rate of 0.001 for 40 episodes.

B ADDITIONAL EXPERIMENTAL RESULTS

Though our problem is a binary classification problem, the target labels CTF are extremely noisy. From Figure 7, we can see that many samples lie around the threshold 6, which is used to decide high and low CTFs in this paper.

B.1 DATASET

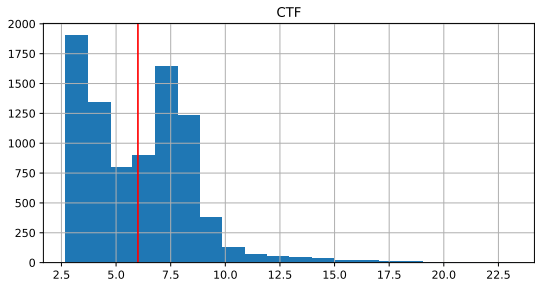


Figure 7: Histogram of the CTF scores over the whole dataset.

B.2 MODEL SELECTION BASED ON ACCURACY

In Table 1, we select model for iteration approaches based on their validation RL rewards. In this section, we compare the model selected by the best accuracy. The results remains the same for most of the cells.

Table 3: A summary of RL rewards and classification accuracy of compared methods. Table (a) shows the average RL rewards and their standard deviation for different methods under 5%, 10%, 20% and 100% of labeled training dataset. Bold text marks the best RL rewards for each row. Table (b) shows the classification accuracy for the perception model. For the iterative methods, we report the results that reaches the highest RL reward over 10 independent runs.

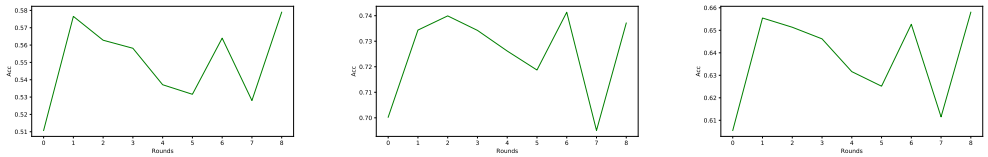
(a) RL rewards

% of labels	*SL*	*FixMatch*	*FixMatch* +iteration	*SSL*^{2}*-RL* 120	*SSL*^{2}*-RL* 480
5%	59.55 $\pm$ 5.4	56.97 $\pm$ 3.2	**62.33** $\pm$ 7.5	61.75 $\pm$ 6.9	61.62 $\pm$ 7.1
10%	50.96 $\pm$ 5.6	58.50 $\pm$ 5.5	59.32 $\pm$ 2.6	64.28 $\pm$ 8.5	**65.73** $\pm$ 7.0
20%	56.76 $\pm$ 7.3	58.98 $\pm$ 3.5	65.77 $\pm$ 4.2	64.29 $\pm$ 8.2	**67.28** $\pm$ 6.3
100%	69.76 $\pm$ 2.1	-	-	-	-

(b) Classification accuracy

% of labels	*SL*	*FixMatch*	*FixMatch* +iteration	*SSL*^{2}*-RL* 120	*SSL*^{2}*-RL* 480
5%	0.5707	0.6229	0.6372	0.646	**0.6451**
10%	0.6188	0.6303	0.653	0.6480	**0.6557**
20%	0.6299	0.6382	0.6396	**0.6502**	0.6479
100%	0.6524	-	-	-	-

B.3 Quality of the pseudo labels



(a) Accuracy of the first half (b) Accuracy of the second half (c) Overall accuracy

Figure 8: Accuracy of the pseudo labels generated by RL trajectories at different rounds for *SSL*^{2}*-RL* 480 for 10% of labeled data.

B.4 ANOVA test

We run ANOVA test on the reduction of sum of squares of CTF scores at different magnification levels.

Table 4: ANOVA test on different magnification levels

Levels	Sum of Square	Cumulative	df	p-value
Grid	5393.9	116737	9	1.1e-16
Square	17167.2	111343	58	1.1e-16
Patch	31570.4	94175	771	1.1e-16
Hole	62604	62604	5997	-

C Proof of Lemma

Proof. We first rewrite the empirical regularization term:

$$\hat{\mathcal{R}}_{\mathcal{B}}(\pi) = \sum_{b=1}^{B} n_b \mathbf{1}(\exists s_1, s_2 \in P_b, (\pi(s_1) - N_C) \times (\pi(s_2) - N_C) < 0).$$

The cumulative penalty term are the total number of switches between partitions. For each partition, whenever $\exists s_1, s_2 \in P_b, (\pi(s_1) - N_C) \times (\pi(s_2) - N_C) < 0$, an extra switch is introduced. Thus we

have the first inequality:

$$\sum_{i=1}^{N_L+N_U} c(S_i, S_{i+1}) \geq B + \sum_{b=1}^{B} \mathbf{1}(\exists s_1, s_2 \in P_b, (\pi(s_1) - N_C) \times (\pi(s_2) - N_C) < 0)$$
$$\geq B + \frac{1}{\min_b n_b} \hat{\mathcal{R}}_{\mathcal{B}}(\pi).$$

The equality can be achieved when π visits the states in a cluster all at once unless some of them have different labels. □