Supplemetary Material

Tiberiu Sosea Cornelia Caragea
University of Illinois Chicago
tsosea2@uic.edu cornelia@uic.edu

A. Additional details of threshold Class C+1

We train the model to output class C+1 on threshold examples. We aim to mimic the margin of mislabeled examples (e.g., an image of a bird labeled as a cat) and intentionally mislabel a random set of unlabeled examples with a new virtual class. By moving these examples in this virtual class $C+1$, we ensure that this class likely contains examples from each of the C classes which will be mislabeled with the inexistent class C+1. For an unlabeled example that now belongs to this virtual class, the model tends to label it in its correct class $c \in C$ through generalization (based on other images that correctly belong to class c), so the difference between the virtual class $C+1$ and the argmax (i.e., probably of c) will be negative, which is similar to the margins of mislabeled examples.

B. Insights into why EMA Margin outperforms EMA Confidence and Entropy

Let x_i be an unlabeled example where our model M fluctuates significantly between iterations. Entropy captures uncertainty through the class probability distribution, but it does not consider the argmax of the prediction. If the class distribution of x_i has one peak but the peak fluctuates between classes from one iteration to another, the average entropy will be low and hence, x_i will be incorrectly included in the training. Confidence measures the magnitude of the probability of predicting a class c and does not penalize too much cases when the distribution has two peaks with fairly similar probabilities (with the other classes having probabilities close to 0). However, unlike entropy, it penalizes changes in argmax across iterations. The margin solves both drawbacks. Let t be the current iteration and c be the argmax at this iteration (hence the *ground truth*; please see line 313 in the main paper). If at some iteration $t' < t$ the predicted class is c', $c' \neq c$, then the difference between the logits corresponding to c and c' is negative since c' corresponds to the argmax at t'. Thus, frequent fluctuations up to iteration t will yield likely negative averaged marings even if the margin for the *ground truth* class c will be large at iteration t but small at iteration t'.

Method	Error Rate
MarginMatch	25.37
FixMatch	38.43
FlexMatch	29.56
OpenCos	31.53
OpenMatch	35.12

Figure 1. Error rates on STL-10 with 4 examples per class.

C. Comparison with Robust SSL Approaches

We carried out an additional experiment to compare our method with OpenMatch [2] and OpenCos [1]. Since both these approaches address the case where the unlabeled data contains a different label space from the labeled data, we only include results on STL-10 using 4 examples per class, since it is the only dataset with these properties. We show these results in Table 1 where we observe that both FlexMatch and our MarginMatch outperform both methods considerably.

References

[1] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. In *ECCV Workshops*, 2021. 1

[2] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25956–25967. Curran Associates, Inc., 2021. 1