SUPPLEMENTARY OF "ON NON-RANDOM MISSING LABELS IN SEMI-SUPERVISED LEARNING"

Xinting Hu[1] **Yulei Niu**[1]* **Chunyan Miao**[1] **Xian-Sheng Hua**[2] **Hanwang Zhang**[1]
[1]Nanyang Technological University, [2]Damo Academy, Alibaba Group
xinting001@e.ntu.edu.sg, yn.yuleiniu@gmail.com
{ascymiao, hanwangzhang}@ntu.edu.sg, xiansheng.hxs@alibaba-inc.com

A APPENDIX

A.1 IMPLEMENTATION DETAILS

As mentioned in Section 5.1, we used almost identical hyper-parameters as FixMatch (Sohn et al., 2020) on CIFAR-10, CIFAR-100, STL-10 and mini-ImageNet. Here, we provide a complete list of hyper-parameters in Table A.1, where xx/xx denote the parameters in FixMatch/Ours and xx is the common parameter.

	Notation	CIFAR-10 iNaturalist-20	CIFAR-100 iNaturalist-50	STL-10	mini-ImageNet
confidence threshold	τ_o	0.95			
unlabel loss weight	λ_u	1			
#unlabeled/#label in batch	η	7			
labeled data batch-size	b	64			
start learning rate	lr	0.03			
momentum	m	0.9			
CAP coefficient	μ	NA/0.99			
CAI coefficient	β	NA/0.5			
weight decay	w	0.0005	0.001/0.0015	0.0005	0.0005

Table A.1: Complete list of hyper-parameters for CIFAR-10, CIFAR-100, STL-10 and mini-ImageNet.

For DARP (Kim et al., 2020), the original implementation needs abundant labeled data as the validation set to estimate the unlabeled data distribution, which is not practical in our case (the minimal number of labeled data of classes is 1). As an alternative, we provide DARP with the ground truth unlabeled data distribution.

A.2 DOUBLE ROBUSTNESS OF THE DR ESTIMATOR

Scenario 1: CAP is correct, *i.e.*, the propensity $p^{(i)}$ successfully reflects the data missing mechanism. We have:

$$\mathcal{L}_{\text{CAI}} + \mathcal{L}_{\text{supp}} = \frac{1}{N} \sum_{i=1,\cdots,N} (1 - \frac{1-m^{(i)}}{p_i}) \mathcal{L}_u(x^{(i)}, q^{(i)}) \, \mathbb{I}(\text{con}(q^{(i)}) > \tau). \tag{A.1}$$

The expectation of it over M is written as:

$$E_M[\mathcal{L}_{\text{CAI}} + \mathcal{L}_{\text{supp}}] = E_M[\sum_{i=1,\cdots,N} (1 - \frac{1-m^{(i)}}{p^{(i)}}) \mathcal{L}_u(x^{(i)}, q^{(i)}) \, \mathbb{I}(\text{con}(q^{(i)}) > \tau)] \tag{A.2}$$

$$= \sum_{i=1,\cdots,N} E_{m^{(i)}}[(1 - \frac{1-m^{(i)}}{p^{(i)}}) \mathcal{L}_u(x^{(i)}, q^{(i)}) \, \mathbb{I}(\text{con}(q^{(i)}) > \tau)]. \tag{A.3}$$

*Now in Columbia University

As $E_{m^{(i)}}[1-m^{(i)}] = p^{(i)}$ is the probability that a data is labeled, we have Eq. A.3 equals 0.

Scenario 2: CAI is correct, *i.e.*, the imputed label plays the same role as the true label:

$$\begin{aligned}\mathcal{L}_{\text{CAP}} + \mathcal{L}_{\text{supp}} &= \frac{1}{N}\sum_{i=1,\cdots,N}\frac{(1-m^{(i)})\mathcal{L}_s(x^{(i)},y^{(i)})}{p^{(i)}} - \frac{1}{N}\sum_{i=1,\cdots,N}(1-m^{(i)})\mathcal{L}_s(x^{(i)},y^{(i)}) \\ &+ \frac{1}{N}\sum_{i=1,\cdots,N}(1-m^{(i)}-\frac{1-m^{(i)}}{p^{(i)}})\mathcal{L}_u(x^{(i)},q^{(i)})\,\mathbb{I}(\text{con}(q^{(i)})>\tau) \\ &= \frac{1}{N}\sum_{i=1,\cdots,N}(\frac{1-m^{(i)}}{p^{(i)}} - (1-m^{(i)})\mathcal{L}_s(x^{(i)},y^{(i)})) \\ &- \frac{1}{N}\sum_{i=1,\cdots,N}(\frac{1-m^{(i)}}{p^{(i)}} - (1-m^{(i)})\mathcal{L}_u(x^{(i)},q^{(i)})\,\mathbb{I}(\text{con}(q^{(i)})>\tau)) \\ &= \sum_{i=1,\cdots,N}(\frac{1-m^{(i)}}{p^{(i)}} - (1-m^{(i)}))(\mathcal{L}_s(x^{(i)},y^{(i)}) - \mathcal{L}_u(x^{(i)},q^{(i)})\,\mathbb{I}(\text{con}(q^{(i)})>\tau)),\end{aligned} \tag{A.4}$$

whose expectation equals to 0 as the second term in summation is expected to be 0 when imputation is ideal.

A.3 Algorithm

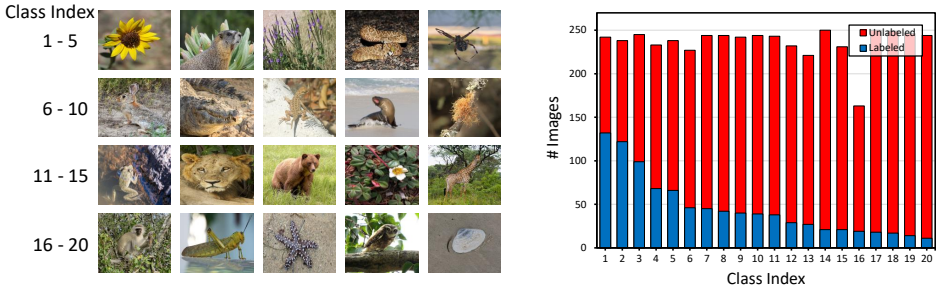
Algorithm 1 Our Class-Aware Doubly Robust Method

Input : D_L, D_U ▷ labeled and unlabeled data
Input : model θ_0 , strong augmentation $\mathcal{A}$, threshold τ_o
Output : θ
Initialize θ_0 randomly, Iteration $i = 0$, $P(Y)$ is uniform
for $i <$ MaxIter **do**
  $\{X_L, Y_L\} \leftarrow D_L$, $\{X_U\} \leftarrow D_U$ ▷ sample a mini-batch
  $\text{Output}_L(\theta) \cup \text{Output}_U(\theta) \leftarrow f_\theta(\{X_L, Y_L\} \cup \{X_U\})$ ▷ Model prediction
  $P(Y|X;\theta) \leftarrow \text{SOFTMAX}(\text{Output}_L(\theta) \cup \text{Output}_U(\theta))$ ▷ softmax probability
  # For supervised labeled data:
  $P(Y), P(Y;\theta) \leftarrow \text{CAP}(\text{Output}_L(\theta) \cup \text{Output}_U(\theta), P(Y))$ ▷ CAP
  $\mathcal{L}_s(X_L, Y_L) \leftarrow \text{CAP-LOSS}(P(Y_L|X_L;\theta), P(Y;\theta))$ ▷ Eq. (5-9)
  # For imputed unlabeled data:
  $Q, \text{con(Q)} \leftarrow \text{MAX}(\text{Output}_U(\theta))$ ▷ imputed label with confidence
  $\tau \leftarrow \tau_o, P(Y), Q$ ▷ Eq. (11)
  $\mathcal{L}_u(X_U) \leftarrow \text{CAI-LOSS}(\text{Output}_U(\theta), \tau)$ ▷ Eq. (13)
  # Model update:
  $\theta \leftarrow \theta - \nabla_\theta \text{DR-LOSS}(\mathcal{L}_s(X_L, Y_L), \mathcal{L}_u(X_U), X_L, P(Y;\theta), \tau)$ ▷ Eq. (14)
end for
Return θ

A.4 More Experiments

A.4.1 More Comparisons of Labeling Dependence on Classes

In this section, we performed experiments with the dependence of labeling on class varying in the intermediate range by increasing the labeling imbalance ration γ from 1 to 200 gradually. As shown in Table A.2, our method can consistently boost the performance of baseline FixMatch under different levels of label dependence. Besides, the improvement is more significant with larger γ, *i.e.*, the



(a) (b)

Figure A.1: Details about our constructed 20 classes subset of the iNaturalist dataset. (a) The example image for each class. (b) Class distribution of the labeled and unlabeled training data.

high dependence of labeling on class. This observation is reasonable since the baseline method gradually fails to handle the data distribution shift between the labeled and unlabeled data in challenging MNAR problems.

Methods	$\gamma = 1$	2	5	10	20	50	100	200
FixMatch	78.54	76.77	74.71	70.53	66.79	59.13	54.78	50.62
Ours	79.02	77.71	76.32	74.37	73.47	70.06	64.47	63.30

Table A.2: Comparison of mean accuracies (%) under different labeling dependence on the class. We alter the imbalance ratio γ of the labeled data in the intermediate range from 1 to 200, where $\gamma = 1$ is the case where the label is missing completely at random. The experiments are conducted on CIFAR-100, and we keep $N_{max} = 200$ across all settings.

A.4.2 Application to iNaturalist

To apply our methods to real-data MNAR occasions, we conducted experiments on the subsets of iNaturalist (Van Horn et al., 2018), a real-world dataset comprised of the natural images and labels collected from a citizen science website[1]. iNaturalist has two popular versions, iNaturalist-2018 for long-tailed recognition[2] and iNaturalist-2021 for nearly balanced data recognition[3]. As iNaturalist-2021 supplements iNaturalist-2018 with abundant additional data in their overlapped classes, these two versions can be naturally used for our MNAR setting in SSL. Specifically, we sampled N classes from the iNaturalist dataset, where the data in iNaturalist-2018 are used as the labeled data and the additionally released data in iNaturalist-2021 as the unlabeled. Figure A.1 depicts the details of our 20-class subset, and the data distribution of 20/50-class subsets are shown in Figure A.1(b) and Figure A.2 separately, where an obvious imbalanced distribution over the labeled data is observed.

	$N = 20$	50
Supervised	25.30	17.09
FixMatch	43.20	47.24
w/ CAP	48.80	**51.32**
w/ CAI	47.50	48.43
w/o CADR	<u>50.80</u>	49.48
w/ CADR	**51.60**	<u>50.14</u>

Table A.3: Comparisons of mean accuracies (%) on the iNaturalist-subsets between the fully-supervised method, FixMatch, and ours. We marked the **best** and <u>second-best</u> accuracies.

To train the dataset, we resize the images to 64×64, remove the overlapped images and sample 50 images per class from iNaturalist2021 for performance evaluation, and train each network with

[1] www.inaturalist.org
[2] https://sites.google.com/view/fgvc5/competitions/inaturalist
[3] https://sites.google.com/view/fgvc8/competitions/inatchallenge2021

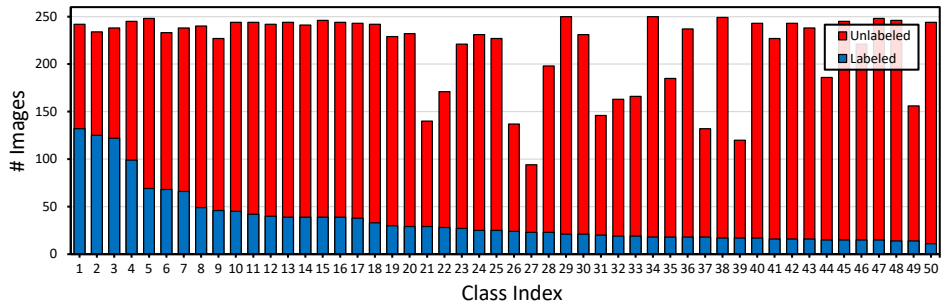


Figure A.2: Class distribution of the labeled and unlabeled training data in our constructed 50 classes subset of the iNaturalist dataset.

Wide ResNet (WRN)-28-2 by 2^{15} iterations. Other hyper-parameters are shown in Table A.1. The performance comparisons are shown in Table A.3. It shows that our proposed methods outperform the baseline FixMatch by large margins, demonstrating the effectiveness of our methods in handling real-world imbalanced labeled data.

A.4.3 More Ablation on Other Baselines

In this section, we perform the ablation experiments on more baseline methods, MixMatch (Berthelot et al., 2019b) and RemixMatch (Berthelot et al., 2019a). As they do not have a threshold in label imputation, CAI is not directly applicable, and we only apply CAP on them. As shown in Table A.4, our method consistently boosts the performance, especially outperforming the baselines by large margins on CIFAR-10 and STL-10.

	CIFAR-10			CIFAR-100		STL-10	mini-ImageNet
Methods	γ=*20*	*50*	*100*	*100*	*200*	*100*	*100*
MixMatch	26.63	31.28	28.02	41.32	42.92	28.31	18.30
w/ CAP	40.34	43.51	45.47	42.45	46.54	34.76	22.09
ReMixMatch	41.84	38.44	38.20	39.71	39.22	39.55	23.50
w/ CAP	51.90	55.03	53.44	40.15	39.40	42.53	23.74

Table A.4: Comparison of mean accuracies (%) with more baselines. We alter the imbalance ratio γ of labeled data and leave the unlabeled data balanced ($\gamma_u = 1$). We keep $N_{max} = \gamma$ so that the least number of labeled data among all the classes is always 1.

REFERENCES

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. In *International Conference on Learning Representations*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019b.

Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist Species Classification and Detection Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.