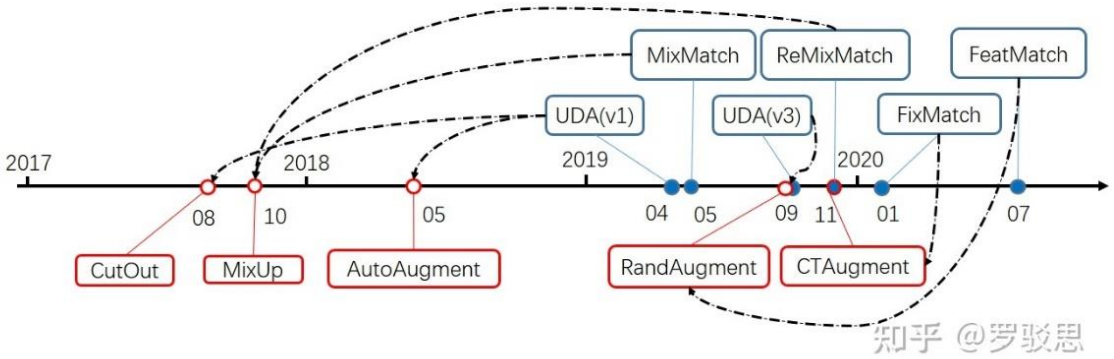
AN UNCERTAINTY-AWARE PSEUDO-LABEL SELECTION FRAMEWORK FOR SEMI-SUPERVISED LEARNIN

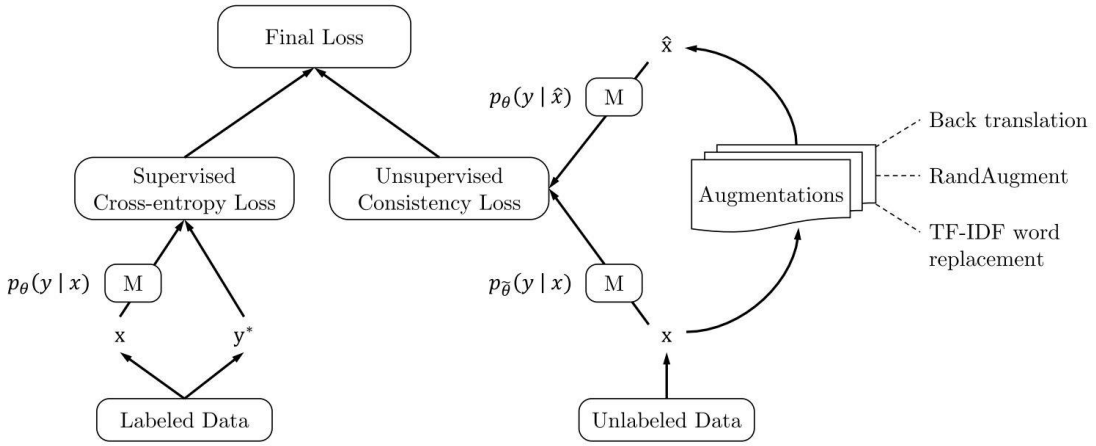
Mamshad Nayeem Rizve[†], **Kevin Duarte**[†], **Yogesh S Rawat**[‡] **& Mubarak Shah**[‡]
Center for Research in Computer Vision
University of Central Florida, Orlando, Florida, USA
[†]{nayeemrizve, kevin_duarte}@knights.ucf.edu
[‡]{yogesh, shah}@crcv.ucf.edu

ICLR 2021

Introduction

Semi-supervised learning (SSL) is mostly dominated by consistency regularization based methods.





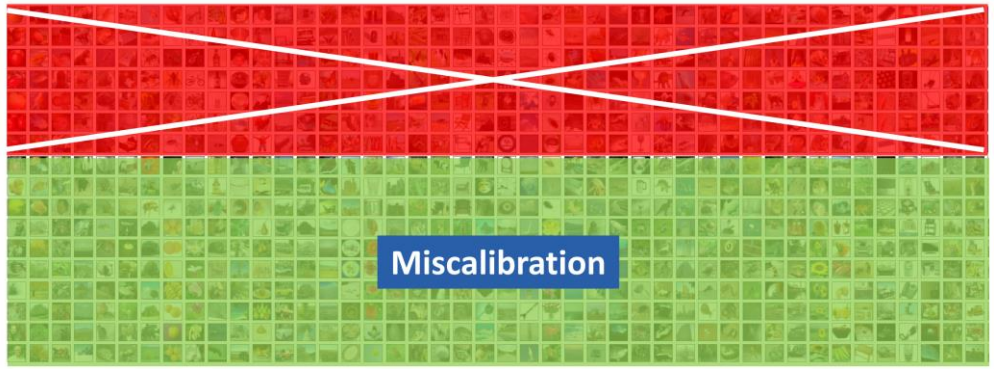
Heavily rely on domain-specific data augmentations, which are not easy to generate for all data modalities.

Overconfidence

So, why pseudo-labeling underperforms in SSL?
the erroneous high confidence predictions from poorly calibrated models.

The ResNet's accuracy is better but not match its confidence.

ResNet 101, Cifar 100
Samples with 80%-85% confidence



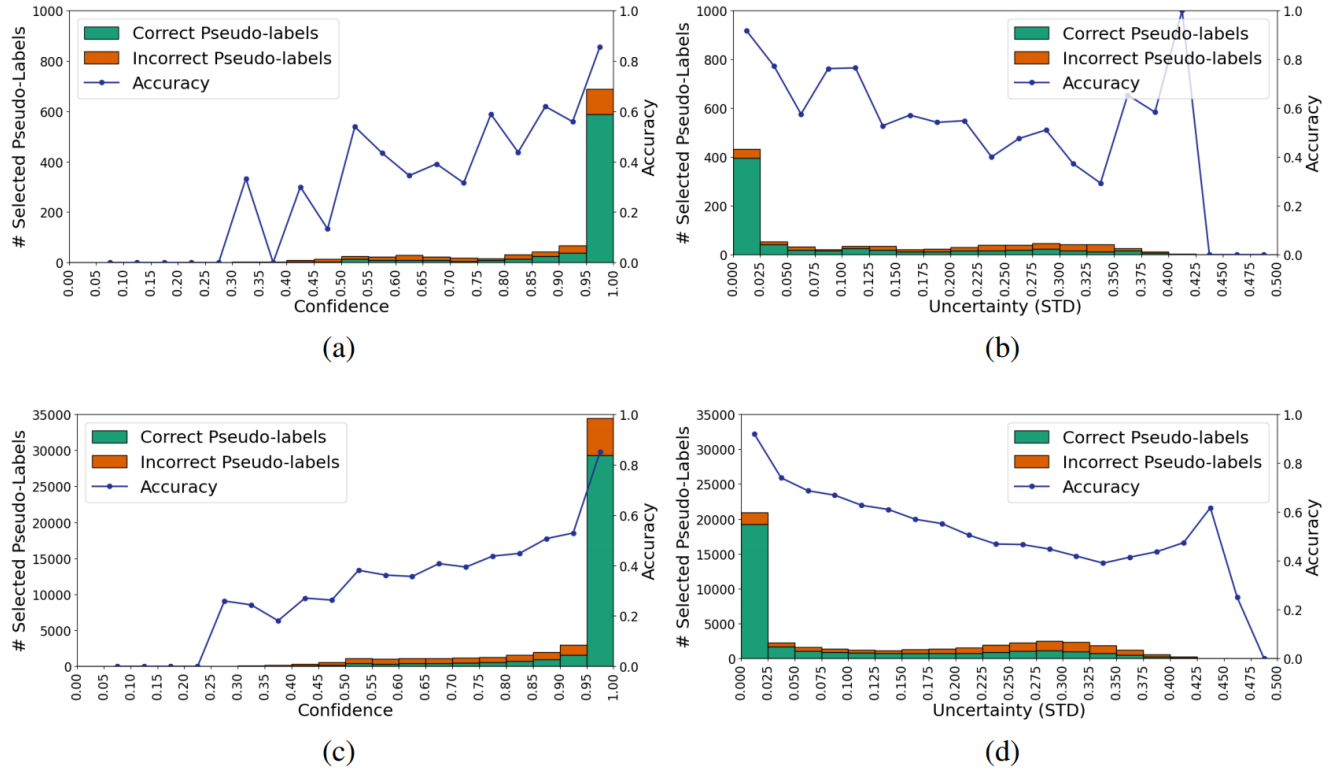


Figure 3: (a) The distribution of correct and incorrect pseudo-labels with respect to the confidence on a 1000 sample CIFAR-10 validation set. (b) The distribution of correct and incorrect pseudo-labels with respect to the uncertainty on a 1000 sample CIFAR-10 validation set. (c) The distribution of correct and incorrect pseudo-labels with respect to the confidence on the set of unlabeled samples. (d) The distribution of correct and incorrect pseudo-labels with respect to the uncertainty on the set of unlabeled samples. Both the full unlabeled set and the small validation have similar skewed confidence and uncertainty distributions.

Calibration of neural networks

The poor calibration of neural networks renders this solution insufficient - in poorly calibrated networks, incorrect predictions can have high confidence scores.

- ECE can be approximated by:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \Big|$$

- Where:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

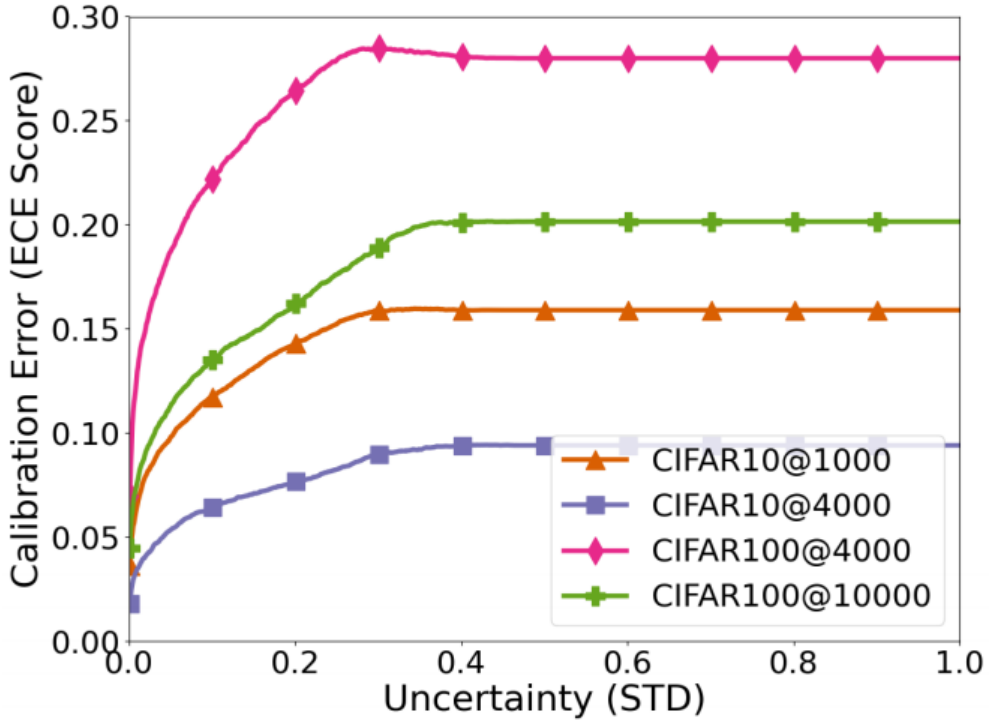
	Avg conf:	0.55	0.74	0.86
—	Accuracy:	0.50	0.67	0.71
	Gap:	2 x 0.05	6 x 0.17	7 x 0.15

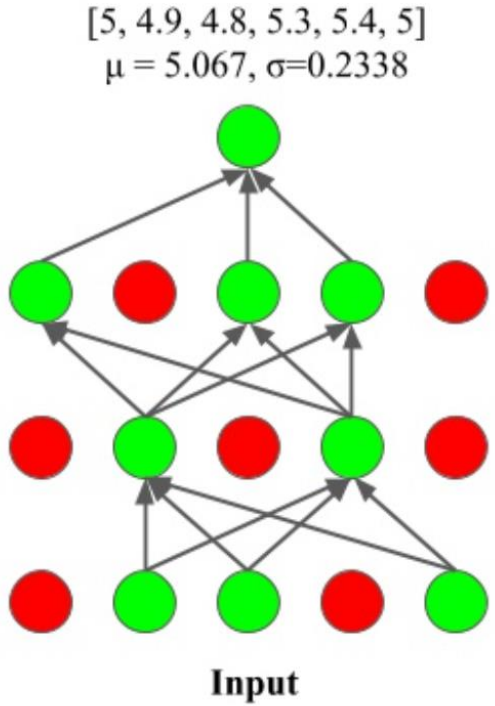
15

Expected Calibrated Error (ECE) = 0.11

Generalizing Pseudo-label Generation

The relationship between the Expected Calibration Error (ECE) score and output prediction uncertainties.





Negative learning

A binary vector representing the selected pseudo-labels in sample

$$g_c^{(i)} = \mathbb{1}\left[p_c^{(i)} \geq \tau_p\right] + \mathbb{1}\left[p_c^{(i)} \leq \tau_n\right],$$

Combine with uncertainty

$$g_c^{(i)} = \mathbb{1}\left[u\left(p_c^{(i)}\right) \leq \kappa_p\right] \mathbb{1}\left[p_c^{(i)} \geq \tau_p\right] + \mathbb{1}\left[u\left(p_c^{(i)}\right) \leq \kappa_n\right] \mathbb{1}\left[p_c^{(i)} \leq \tau_n\right]$$

Loss Function

$$L_{\mathrm{BCE}}\left(\tilde{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \boldsymbol{g}^{(i)}\right) = -\frac{1}{s^{(i)}} \sum_{c=1}^{C} g_c^{(i)} \left[\tilde{y}_c^{(i)} \log\left(\hat{y}_c^{(i)}\right) + \left(1 - \tilde{y}_c^{(i)}\right) \log\left(1 - \hat{y}_c^{(i)}\right)\right].$$

Algorithm

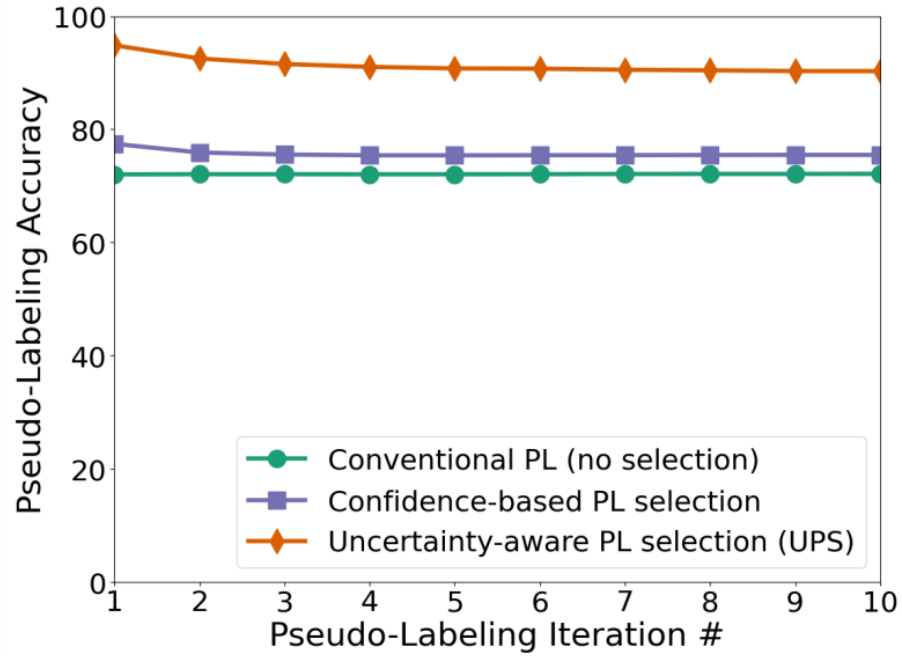
B UPS Training Procedure

The training procedure for our proposed UPS framework is described in Algorithm 1.

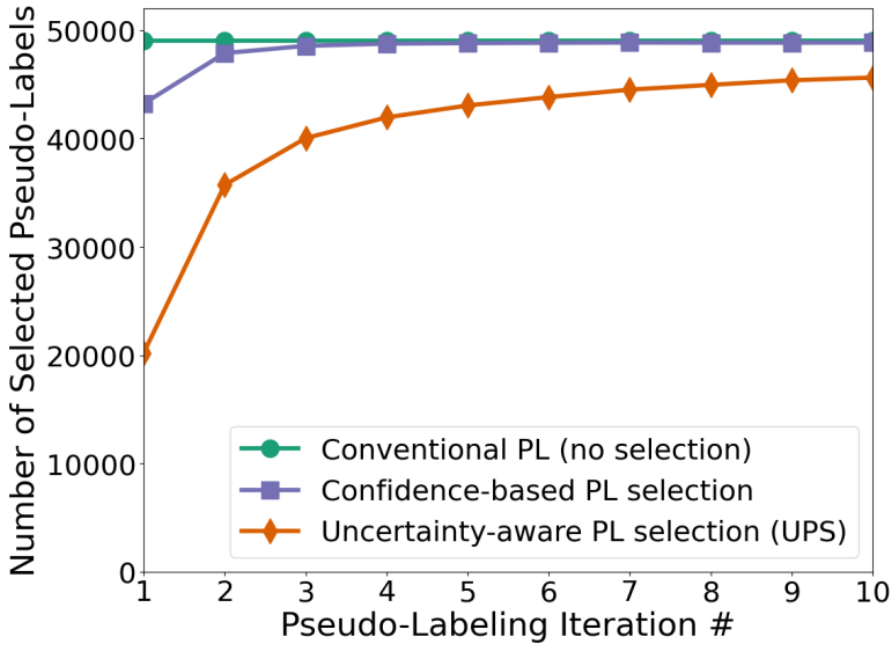
Algorithm 1 The proposed method takes a set of labeled data, D_L, and a set of unlabeled data, D_U, and returns a trained model, f_θ, using samples from both D_L and D_U

Train a network, $f_{\theta,0}$, using the samples from D_L.
for i = 1..MaxIterations **do** ▷ Repeats until convergence
    Pseudo-label D_U using $f_{\theta,i-1}$ ▷ Equation 1
    $D_{\text{selected}} \leftarrow$ Select pseudo-labels using UPS ▷ Equation 5
    $\tilde{D} \leftarrow D_L \cup D_{\text{selected}}$
    Initialize new network $f_{\theta,i}$
    Train $f_{\theta,i}$ using the samples from $\tilde{D}$. ▷ Using cross-entropy loss or losses in Equations 4-5
    $f_\theta \leftarrow f_{\theta,i}$
return f_θ

Experiments



(b)



(c)

Experiments

Table 1: Error rate (%) on the CIFAR-10 and CIFAR-100 test set. Methods with † are pseudo-labeling based, whereas others are consistency regularization methods.

Method	CIFAR-10		CIFAR-100	
	1000 labels	4000 labels	4000 labels	10000 labels
DeepLP†	22.02 ± 0.88	12.69 ± 0.29	46.20 ± 0.76	38.43 ± 1.88
TSSDL†	21.13 ± 1.17	10.90 ± 0.23	-	-
MT	19.04 ± 0.51	11.41 ± 0.25	45.36 ± 0.49	36.08 ± 0.51
MT + DeepLP	16.93 ± 0.70	10.61 ± 0.28	43.73 ± 0.20	35.92 ± 0.47
ICT	15.48 ± 0.78	7.29 ± 0.02	-	-
DualStudent	14.17 ± 0.38	8.89 ± 0.09	-	32.77 ± 0.24
R2-D2	-	-	-	32.87 ± 0.51
MixMatch	-	6.84	-	-
UPS†	$\mathbf{8.18 \pm 0.15}$	$\mathbf{6.39 \pm 0.02}$	$\mathbf{40.77 \pm 0.10}$	$\mathbf{32.00 \pm 0.49}$

Table 2: Error rate (%) on CIFAR-10 with different backbones Wide ResNet-28-2 (WRN) and Shake-Shake (S-S).

Method	Backbone	Labels	
		1000	4000
MixMatch	WRN	7.75	6.24
MixMatch	S-S	-	4.95
ReMixMatch	WRN	5.73	5.14
TC-SSL	WRN	6.15	5.07
R2-D2	S-S	-	5.72
UPS	WRN	7.95	6.42
UPS	S-S	-	4.86

Table 3: Accuracy (%) on the UCF-101 test set. Methods with * use scores reported in (Jing et al., 2020).

Method	20% labeled	50% labeled
Supervised	33.5	45.6
MT*	36.3	45.8
PL*	37.0	47.5
S4L*	37.7	47.9
UPS	**39.4**	**50.2**

Table 4: mAP scores on the Pascal VOC2007 test set.

Method	10% labeled	20% labeled
Supervised	18.36 ± 0.65	28.84 ± 1.68
PL	27.44 ± 0.55	34.84 ± 1.88
MixMatch	29.57 ± 0.78	37.02 ± 0.97
MT	32.55 ± 1.48	39.62 ± 1.66
UPS	$\mathbf{34.22 \pm 0.79}$	$\mathbf{40.34 \pm 0.08}$

Ablation study

Table 5: Ablation Study on CIFAR-10 dataset (Error Rate (%)). UPS with no uncertainty-aware (UA) selection, selects using only confidence-based criteria.

Method	1000 labels	4000 labels
Supervised	27.66	16.65
UPS, no selection	22.60	12.94
UPS, no UA	16.50	10.02
UPS, no UA (Cal.)	13.68	8.09
UPS, no NL	9.46	6.64
UPS, full method	8.14	6.36

Table 6: Comparison of methods for uncertainty estimation on CIFAR-10 (1000 labels) (Error Rate (%))

Method	1000 labels	4000 labels
MC-Dropout	8.14	6.36
MC-SpatialDropout	8.28	6.60
MC-DropBlock	9.76	7.50
DataAug	8.28	6.72

Table 12: Performance on the CIFAR-10 and Pascal VOC2007 test sets.

Method	CIFAR-10 (accuracy)		Pascal VOC2007 (mAP)	
	1000 labels	4000 labels	10% labeled	20% labeled
UPS, with class balance	91.86	93.64	34.72	40.33
UPS, without class balance	88.77	93.14	31.88	40.06

Generalizing Pseudo-label Generation

Learning from Crowds by Modeling Common Confusions

Zhendong Chu, Jing Ma, Hongning Wang
Department of Computer Science
University of Virginia
{zc9uy, jm3mr, hw5x}@virginia.edu

AAAI 2021

Introduction

The annotation quality of annotators varies considerably

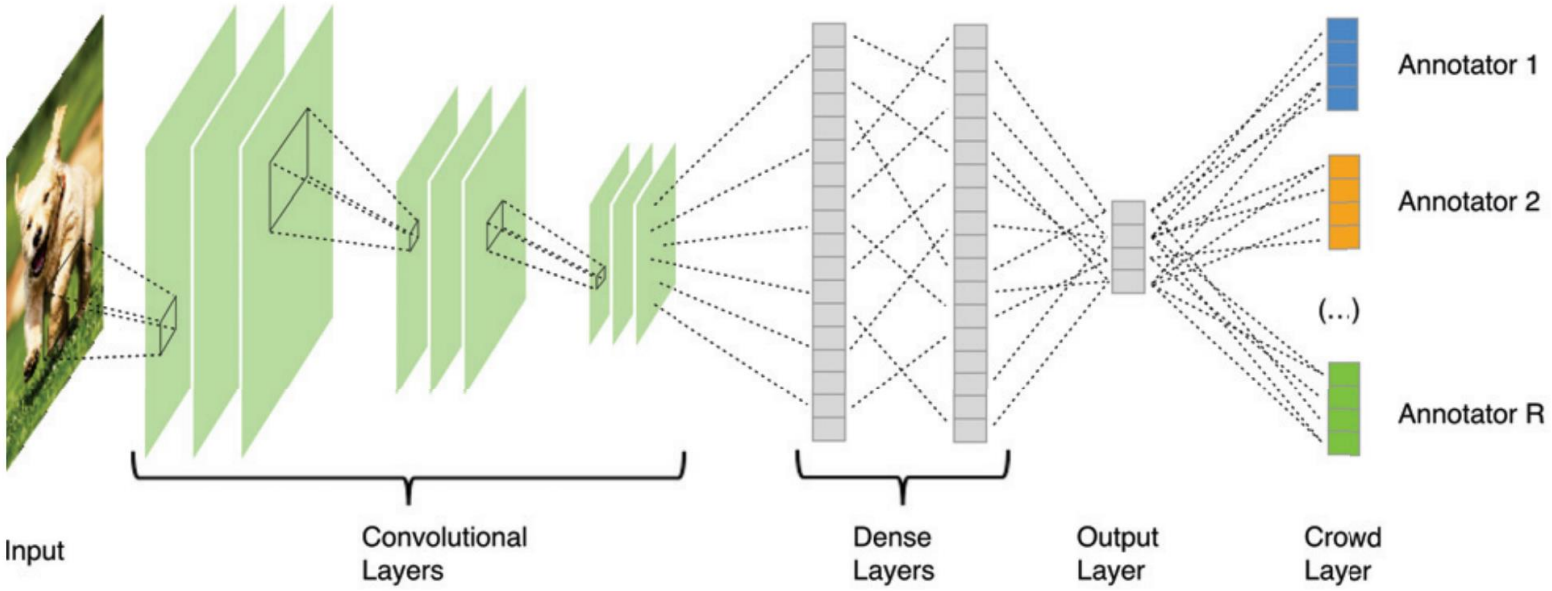
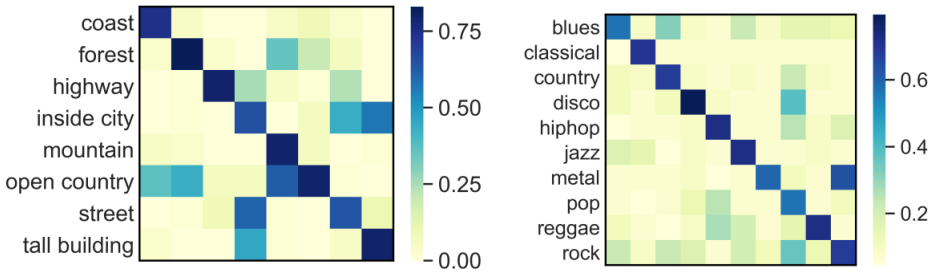


Figure 1: Bottleneck structure for a CNN for classification with 4 classes and R annotators.

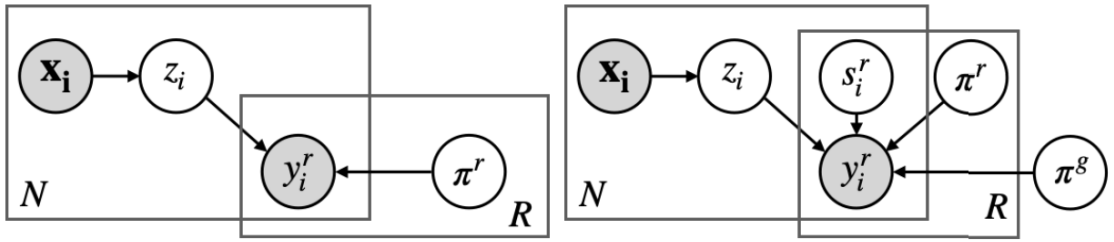
Introduction

The majority of annotators are not necessarily correct, as their mistakes are no longer independent.



(a) LabelMe (b) Music

Figure 1: Analysis of commonly made mistakes across annotators on two real-world crowdsourcing datasets. The value of each entry in the heatmap denotes the percentage of annotators with this confusion pair (e.g., mistakenly label *street* as *inside city* on LabelMe dataset).



(a) DS model. (b) Common noise model.

Figure 2: Graphical model presentations of DS model and our common noise model.

Model

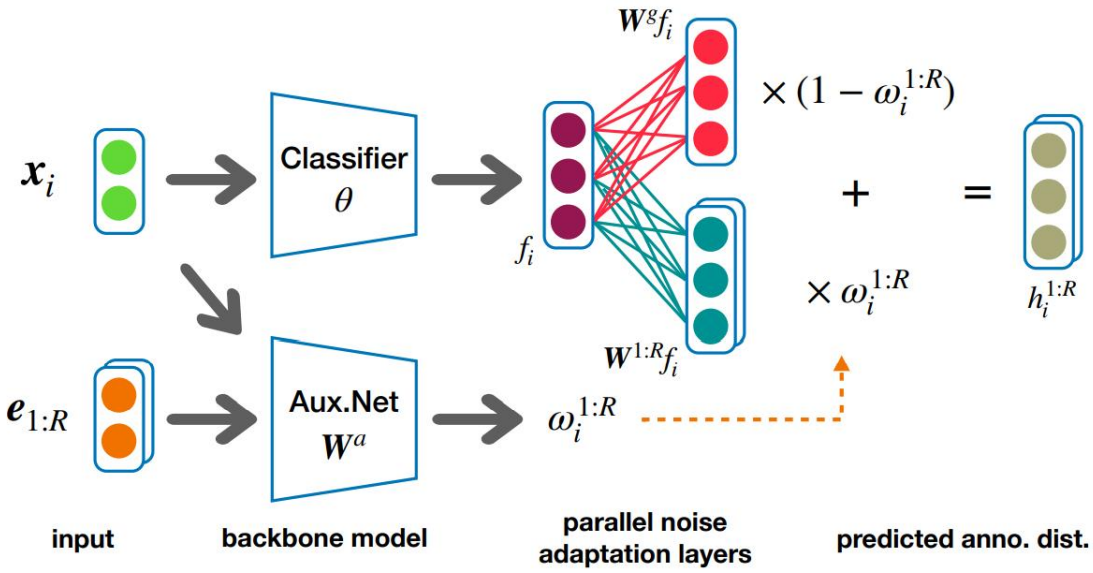


Figure 3: Overview of our framework for classification with 3 classes and R annotators.

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{W}^g, \boldsymbol{W}^{1:R}, \boldsymbol{W}^a) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{j=1}^{C} y_{ij}^r \log p_j(\hat{y}_i^r|\boldsymbol{x}_i) - \lambda \sum_{r=1}^{R} \|\boldsymbol{W}^g - \boldsymbol{W}^r\|_2$$

$$p(\hat{y}_i^r|\boldsymbol{x}_i) = \omega_i^r \, p_{\boldsymbol{W}^g}(\hat{y}_i^r|f(\boldsymbol{x}_i)) + (1 - \omega_i^r)\, p_{\boldsymbol{W}^r}(\hat{y}_i^r|f(\boldsymbol{x}_i)).$$

$$\boldsymbol{v}_i = \boldsymbol{W}_v \boldsymbol{x}_i + b_v, \boldsymbol{u}_r = \boldsymbol{W}_u \boldsymbol{e}_r + b_u,$$

$$\omega_i^r = \sigma(\boldsymbol{u}_r^\top \boldsymbol{v}_i).$$

Experiments

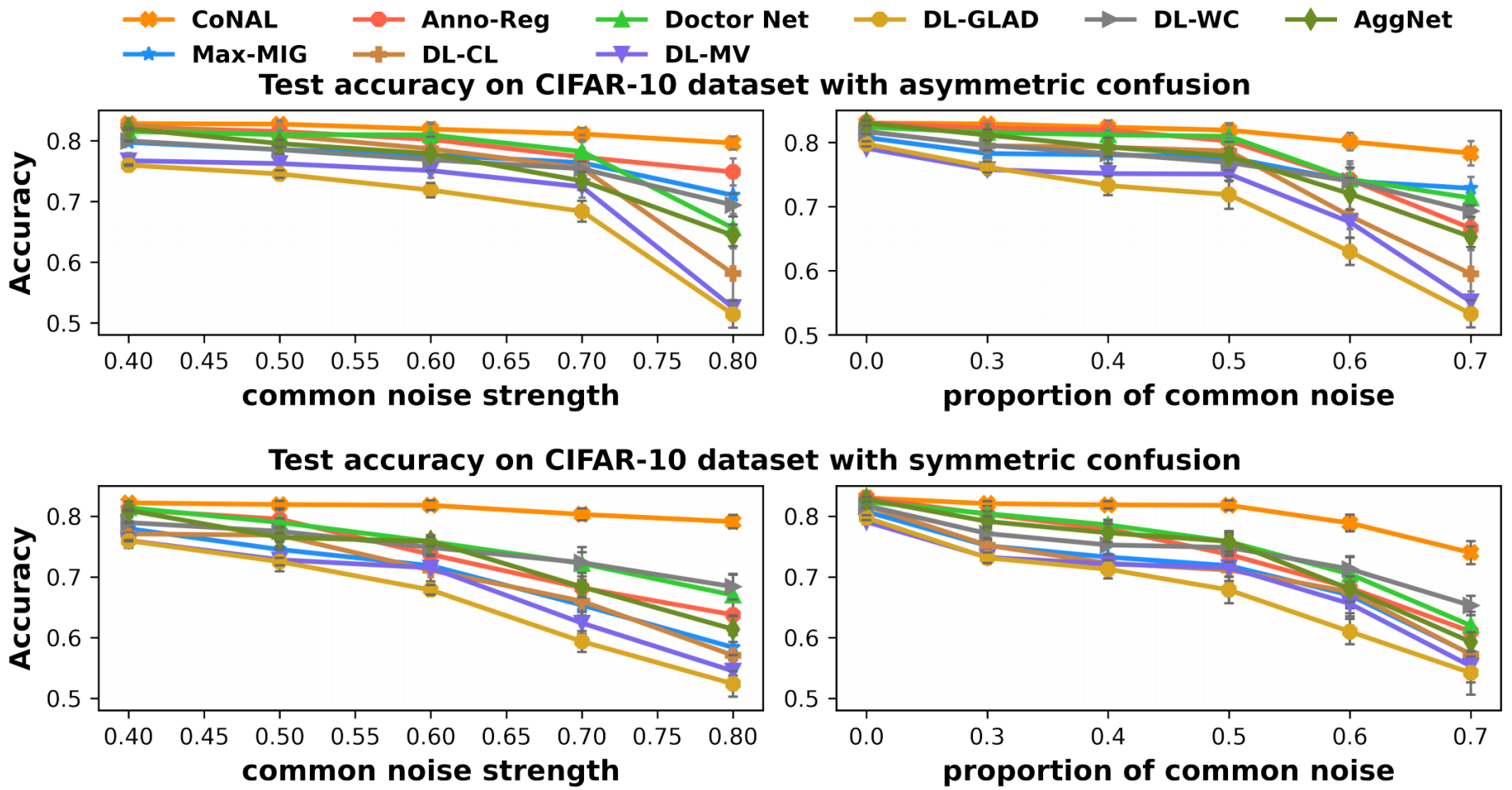


Figure 4: Results on CIFAR-10 dataset.

Experiments

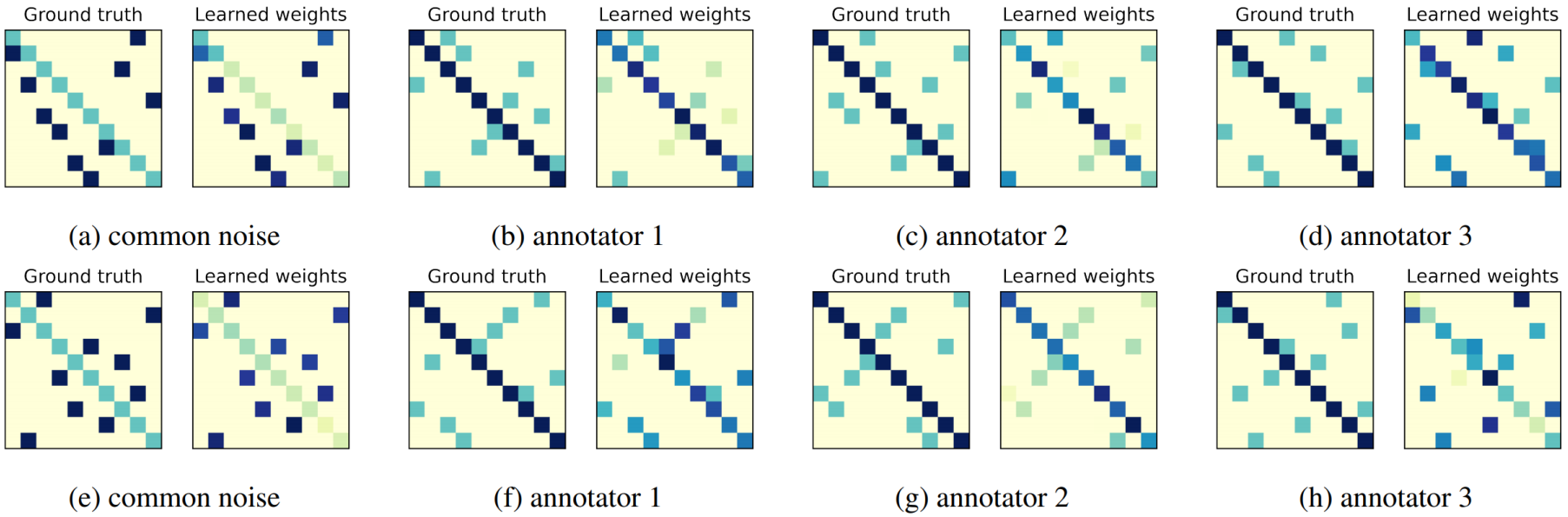


Figure 5: Comparison between ground truth confusion matrices and learned ones on CIFAR-10 dataset. The top row is the result of asymmetric common noise. The bottom row is the result of symmetric common noise.

	DL-MV	DL-CL	Doctor Net	Anno-Reg	Max-MIG	DL-GLAD	DL-WC	AggNet	CoNAL
LabelMe	79.83±0.34	83.27±0.52	82.12±0.43	82.77±0.48	85.33±0.61	83.12±0.34	82.74±0.33	84.75±0.27	**87.12**±0.55
Music	72.53±0.41	81.46±0.53	76.58±0.47	79.12±0.36	81.37±0.33	77.82±0.37	75.76±0.24	81.92±0.41	**84.06**±0.42

Table 1: Test accuracy on two real-world crowdsourcing datasets.