Bootstrapping the Relationship Between Images and Their Clean and Noisy Labels

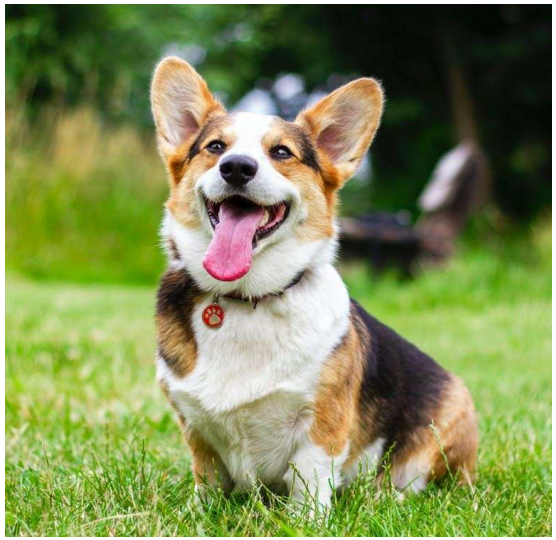
Brandon Smart
Australian Institute for Machine Learning
University of Adelaide, Australia
a1743623@adelaide.edu.au

Gustavo Carneiro
Centre for Vision, Speech and Signal Processing
University of Surrey, United Kingdom
g.carneiro@surrey.ac.uk

WACV 2023

Background

Noisy labels —— In multi-class classification task



Dog (clean)

↓

Cat (noisy)

Noisy type

- symmetric label noise
- ∧
- asymmetric label noise
- ∧
- instance-dependent noise

Background

How to train with noisy labels ?

Recent methods

- train with noisy-labelled data —— **overfit the incorrectly labelled samples**
- sample re-labelling mechanism —— **overlook the *relationship* between images, clean labels and noisy labels**
- model this ***relationship*** —— **failed to achieve SOTA results**

Motivation

Model this *relationship*

The relationship actually helps with **improving robustness**.

However …

①They need both a clean set and noisy set.

Furthermore…

②They require distillation to a more standard model for evaluation.

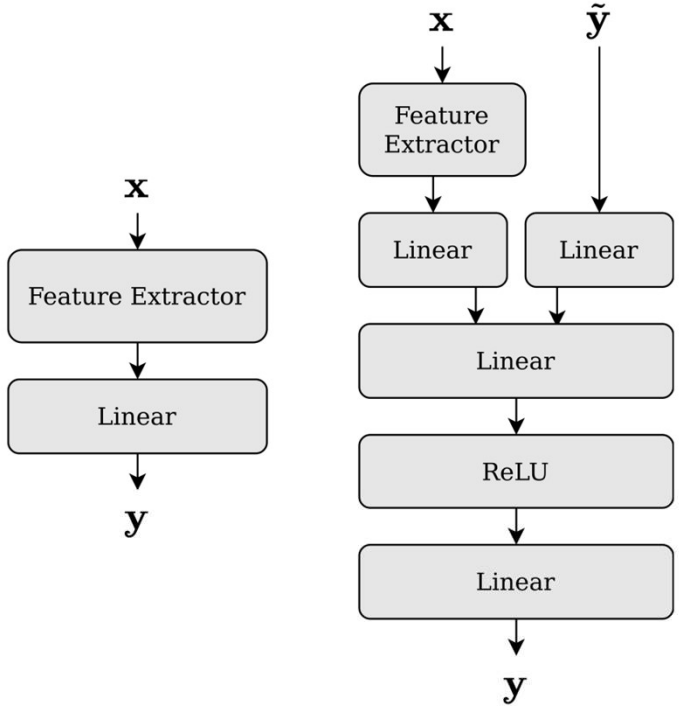


Figure 1. On the left, we show a 'normal' deep neural network model used for noisy label learning tasks. On the right, we present a 'modified' model that can learn the relationship between images $\mathbf{x}$, noisy labels $\tilde{\mathbf{y}}$ and clean labels $\mathbf{y}$, similar to those used by methods that have access to a clean set of data [16, 20, 50].

Method

3-stage process： Bootstrapping， Semi-supervised Learning， Final training

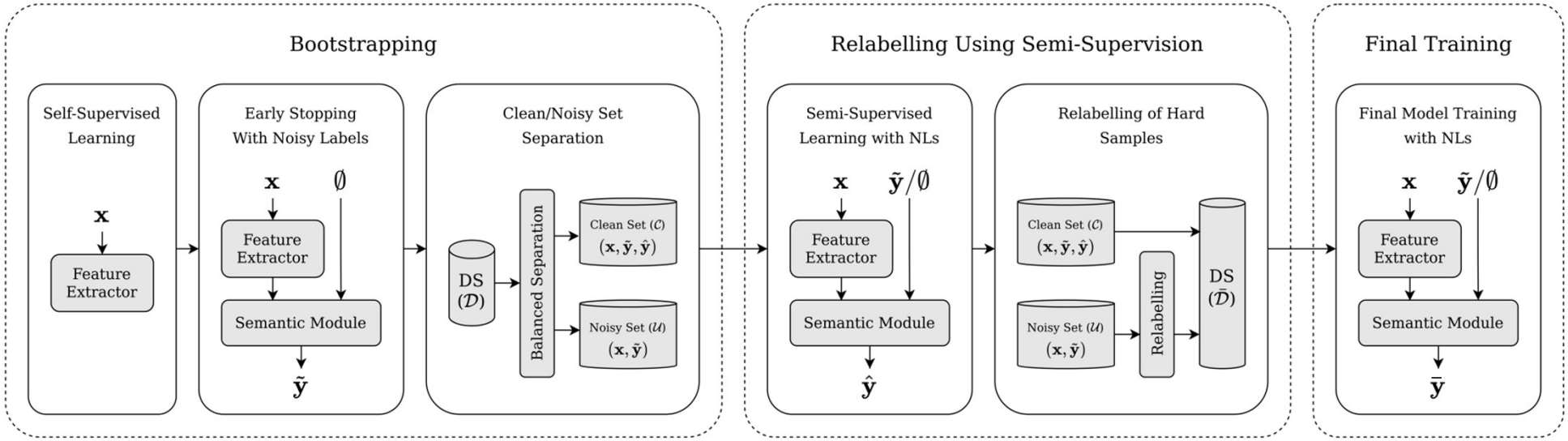
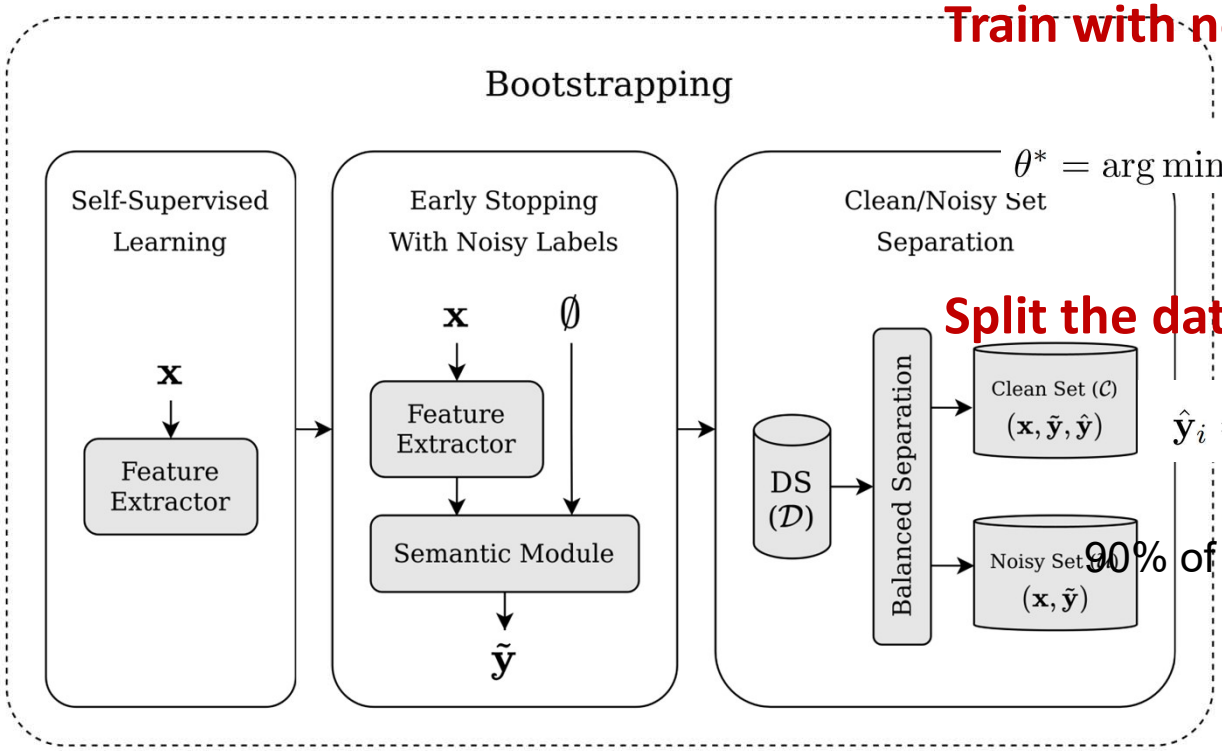


Figure 2. Our proposed algorithm. The bootstrapping stage consists of self-supervised pre-training, followed by early-stopping classification training to identify a small set of confidently relabelled samples (i.e., clean labelled samples). This set is then used to learn the relationship between images, noisy labels and clean labels during semi-supervised learning, and at the end of this stage, we relabel the samples classified as noisy during bootstrapping. The last stage is the final training that trains the classifier using the clean and noisy samples identified in the SSL stage.

Method

1st-stage process： Bootstrapping

Aim to: split data into a **clean** set and a **noisy** set

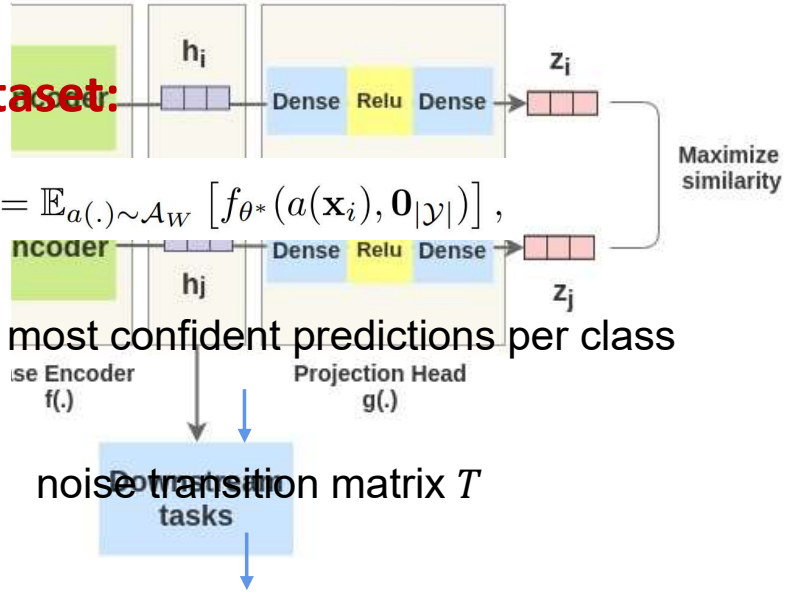


Train with noisy labels:

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \mathbb{E}_{a(.) \sim \mathcal{A}_S} \left[\ell_{CE}(\tilde{\mathbf{y}}_i, f_\theta(a(\mathbf{x}_i), \mathbf{0}_{|\mathcal{Y}|})) \right],$$

Split the dataset:

$$\hat{\mathbf{y}}_i = \mathbb{E}_{a(.) \sim \mathcal{A}_W} \left[f_{\theta^*}(a(\mathbf{x}_i), \mathbf{0}_{|\mathcal{Y}|}) \right],$$



90% of most confident predictions per class

noise transition matrix T

select the K × |Y| × T_{ij} most confident samples

Method

2nd-stage process： Semi-supervised Learning Aim to: re-label the **hard** samples (Noisy Set)

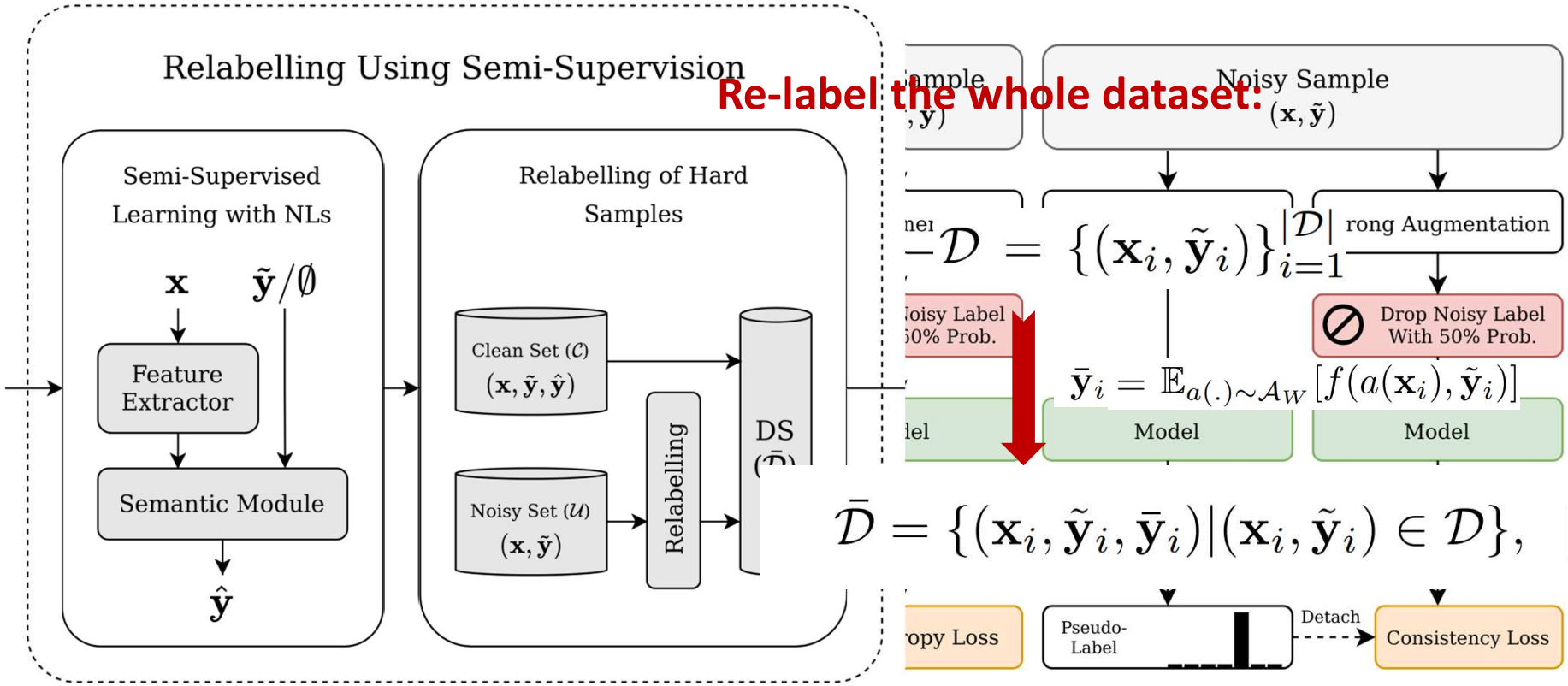
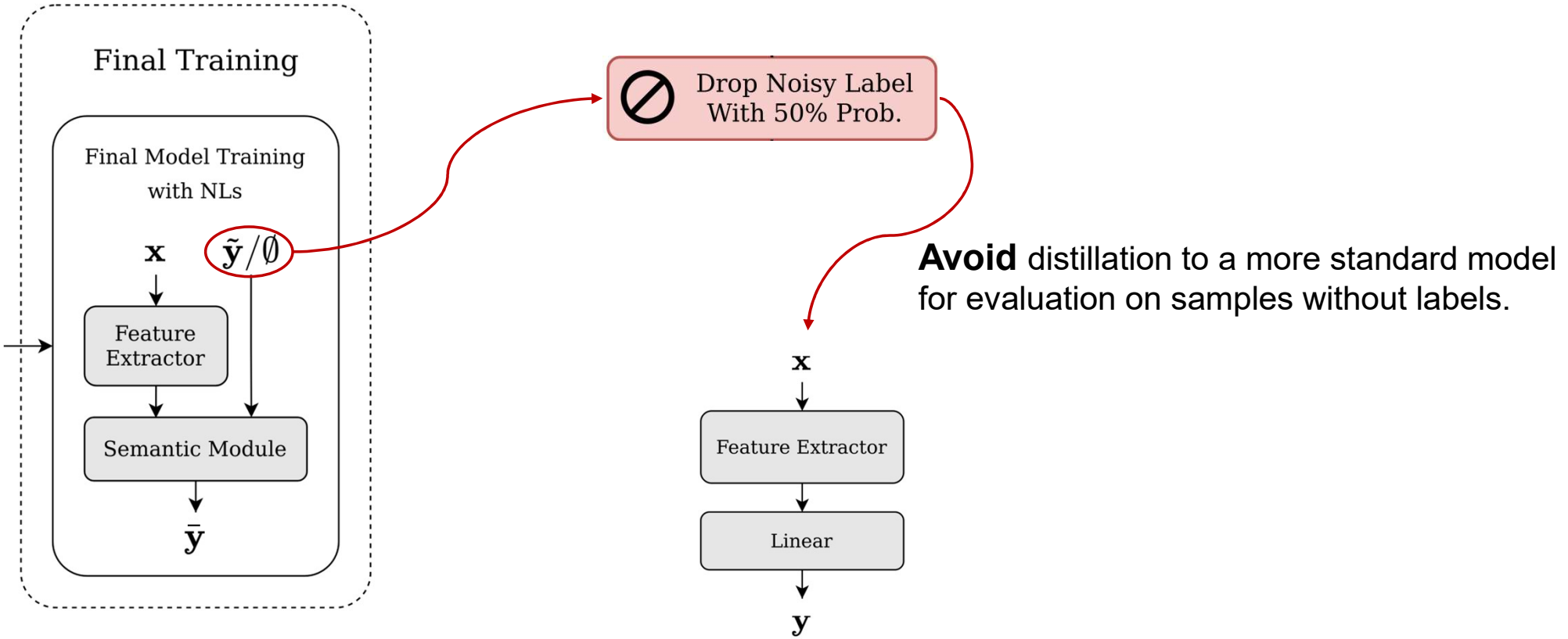


Figure 3. The noisy-label FixMatch algorithm.

Method

3rd-stage process： Final training

Aim to: train a model with the dataset re-labelled



Avoid distillation to a more standard model for evaluation on samples without labels.

Experiment

Comparation to SOTA

Method	Top-1
Cross Entropy	79.4
SELFIE [46]	81.8
PLC [62]	83.4
NCT [42]	84.1
Ours (Normal Model)	85.84
+ *Test-Time Aug.*	*86.98*
Ours (Modified Model)	**88.48**
+ *Test-Time Aug.*	*89.38*

Table 1. Test accuracy (%) for Animal10N. Top methods within 1% in **bold** (Results other than ours are from [42]).

Method	Webvision		ILSVRC2012	
	Top-1	Top-5	Top-1	Top-5
ELR [32]	76.26	91.26	68.71	87.84
ELR+ [32]	77.78	91.68	70.29	89.76
PropMix [10]	78.84	90.56	/	/
NGC [53]	79.16	91.84	74.44	91.04
FaMuS [58]	79.40	**92.80**	**77.00**	**92.76**
RRL [30]	76.30	91.50	73.30	91.20
Ours (Normal Model)	**80.24**	90.84	76.44	90.00
+ *Test-Time Aug.*	*81.52*	*92.52*	*78.36*	*91.80*
Ours (Modified Model)	**80.88**	**92.76**	75.96	**92.20**
+ *Test-Time Aug.*	*83.16*	*94.28*	*79.64*	*94.20*

Table 2. Test accuracy (%) for Webvision. Top methods within 0.5% in **bold**

Comparation to SOTA

Dataset	CIFAR-10			CIFAR-100		
Noise Type	Type-I 35%	Type-II 35%	Type-III 35%	Type-I 35%	Type-II - 35%	Type-III - 35%
Cross-Entropy	78.11	76.65	76.89	57.68	57.83	56.07
PLC [63]	82.80	81.54	81.50	60.01	63.68	63.68
Ours (Normal Model)	**94.06**	**93.25**	**93.35**	65.87	65.80	66.36
+ *Test-Time Aug.*	*94.72*	*93.79*	*93.97*	*66.83*	*66.48*	*67.42*
Ours (Modified Model)	**94.00**	**93.76**	**94.23**	**68.25**	**68.14**	**68.22**
+ *Test-Time Aug.*	*94.39*	*94.19*	*94.23*	*70.13*	*69.35*	*70.13*

Table 3. Test accuracy (%) for Polynomial Margin Diminishing Noise [62]. Top methods are in **bold**.

Data set	CIFAR-10			CIFAR-100		
Method/ Noise Ratio	DenseNet (32%)	ResNet (38%)	VGG (34%)	DenseNet (34%)	ResNet (37%)	VGG (37%)
D2L + RoG [27]	68.57	60.25	59.94	31.67	39.92	45.42
CE + RoG [27]	68.33	64.15	70.04	61.14	53.09	53.64
PropMix [10]	84.25	82.51	85.74	60.98	58.44	60.01
Ours (Normal Model)	**93.26**	**92.05**	**93.29**	62.47	**64.91**	**64.98**
+ *Test-Time Aug.*	*93.87*	*92.66*	*93.86*	*63.40*	*65.74*	*66.10*
Ours (Modified Model)	89.46	90.97	89.77	**63.68**	63.09	63.89
+ *Test-Time Aug.*	90.25	*91.85*	*90.42*	*65.15*	*64.70*	*65.30*

Table 4. Test accuracy (%) for the RoG label noise benchmark [27], where baseline results are from [27]. Top methods are in **bold**.

Experiment

Comparation to SOTA

Dataset	CIFAR-10					CIFAR-100			
Noise type	Sym.				Asym.	Sym.			
Method / Noise Ratio	20%	50%	80%	90%	40%	20%	50%	80%	90%
Cross-Entropy [28]	82.7	57.9	26.1	16.8	72.3	61.8	37.3	8.8	3.5
ELR [32]	**95.8**	**94.8**	93.3	78.7	93.0	77.6	73.6	60.8	33.4
DivideMix [28]	**95.7**	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0
AugDesc [34]	**96.3**	**95.4**	**93.8**	91.9	94.6	**79.5**	**77.2**	66.4	41.2
ContrastToDivide [66]	**96.4**	**95.3**	**94.4**	**93.6**	93.5	**78.7**	**76.4**	**67.8**	**58.7**
PropMix [10]	**96.09**	**95.53**	**93.77**	**93.20**	94.64	76.99	73.71	66.75	**58.32**
Ours (Normal Model)	95.04	**95.13**	**94.51**	91.72	94.91	69.02	68.87	64.09	55.91
+ *Test-Time Aug.*	*95.47*	*95.39*	*94.90*	*92.32*	*95.14*	*70.17*	*69.57*	*65.10*	*56.81*
Ours (Modified Model)	**95.99**	**95.59**	**94.48**	**93.52**	**95.85**	75.09	70.86	57.03	39.95
+ *Test-Time Aug.*	*96.75*	*96.16*	*94.98*	*93.93*	*96.42*	*76.65*	*72.95*	*57.83*	*40.31*

Table 5. Test accuracy (%) for all competing methods on CIFAR-10 and CIFAR-100 under symmetric and asymmetric noises. Results from related approaches are as presented in [28] and [53]. Top methods within 1% are in **bold**.

Predictions with Noisy Labels

Test Set	Accuracy
Without Noisy Labels	95.85
With Noisy Labels	97.59

Table 6. Accuracy of our model tested with and without noisy labels on CIFAR10 Asym. 40% noise. Noisy labels are generated for training and testing samples using the same procedure.

Noisy Label	Prediction (Confidence)
-	Dog (90.27%)
Airplane	Bird (65.39%)
Automobile	Automobile (62.75%)
Bird	Bird (98.30%)
Cat	Dog (94.51%)
Deer	Deer (96.06%)
Dog	Dog (96.29%)
Frog	Frog (96.49%)
Horse	Horse (75.84%)
Ship	Ship (95.46%)
Truck	Truck (96.88%)

Table 7. Predictions made by our model with different noisy labels for a testing sample (showing a dog) in Asym 0.4 noise for CIFAR10. '-' represents using a null label in place of a noisy label.

Ablations and Training Time

Training Stage	Accuracy
After Bootstrapping	91.41
After Semi-Supervised Learning	94.98
After Final Training	95.85

Table 8. Model accuracy after each stage of training on CIFAR10 Asym. 40% noise.

		Evaluation Aug.		
		None	Weak	Strong
Train	None	579	361	456
	Weak	265	56	300
	Strong	28	**21**	31

Table 9. Effect of different training/testing augmentations on the number of errors in a clean set of 10,000 samples selected after bootstrapping. Test performed on CIFAR10 Asym. 40% noise.

Training Strategy	No. of Errors
No Self-Supervision	369
With Self-Supervision	21

Table 10. Effect of self-supervision on the number of errors in a clean set of 10,000 samples selected after bootstrapping. Test performed on CIFAR10 Asym. 40% noise

Thanks