A. Algorithm of the proposed method RETINA

We provide the details of our proposed method RETINA in Algorithm 1.

Algorithm 1 The training of RETINA

procedure MULTI-TEACHER STUDENT TRAINING($\{\mathcal{D}^{(m)}\}_{m=1}^{M}, \lambda_u, n_{\text{warm\_up}}, n_{\text{e}}$)

  ▷ $\mathcal{D}^{(m)}$: *the noisy label dataset of annotator* m ◁

  ▷ λ_u: *a hyper-parameter that weights the loss of noisy label samples* ◁

  ▷ $n_{\text{warm\_up}}$: *the number of warmup epochs* ◁

  ▷ n_{epoch}: *the number of training epochs* ◁

  initialize M model parameters: $\{\theta^{(m)}\}_{m=1}^{M}$

  warm-up on noisy datasets: $\theta^{(m)} \leftarrow \text{WARM-UP}(\mathcal{D}^{(m)}, \theta^{(m)}, n_{\text{warm\_up}}), \forall m \in \{1, \ldots, M\}$

  for epoch $= n_{\text{warm\_up\_epoch}} + 1 : n_{\text{epoch}}$ **do**

    for $m = 1 : M$ **do**

      $f_{\theta^{(m)}} \sim_{\text{w/o}} \mathcal{F}_s(\{f_{\theta^{(m)}}\}_{m=1}^{M}, \{\mathcal{D}^{(m)}\}_{m=1}^{M}$ ▷ *Sample a student without replacement*

      $f_{\theta^{(n)}} \sim \mathcal{F}_t(\{f_{\theta^{(n)}}\}_{n=1}^{M}, \{\mathcal{D}^{(m)}\}_{m=1}^{M}), \text{student} = f_{\theta^{(m)}}, n \neq m)$ ▷ *Select the teacher*

      $\mathcal{D}_{\text{clean}}^{(m)}, \mathcal{D}_{\text{noisy}}^{(m)} \leftarrow \text{SAMPLE-SELECTION}(f_{\theta^{(n)}}, \mathcal{D}^{(m)})$ ▷ *Eq.* (6)

      $\mathsf{L} = \ell_{\text{CLEAN}}(\mathcal{D}_{\text{clean}}^{(m)}, \theta^{(m)}) + \lambda_u \ell_{\text{NOISE}}(\mathcal{D}_{\text{noisy}}^{(m)}, \theta^{(m)}) + \lambda_r \ell_{\text{REG}}$ ▷ *loss to train the student model*

      $\theta^{(m)} \leftarrow \mathsf{SGD}(\mathsf{L}, \theta^{(m)})$ ▷ *train student model and update model parameters*

  return $\{\theta^{(m)}\}_{m=1}^{M}$

B. Detailed experiment setting

Noise ratios in different datasets Tab. 3 presents the noise rates (in percentage) for both individual annotators and their aggregated majority vote labels across several datasets evaluated in our experiments:

	Noise rates of individual annotators and majority vote labels (%)								
№ annotator	a1	a2	a3	a4	a5	m1∼2	m1∼3	m1∼4	m1∼5
CIFAR100-IDN30	30.07	30.11	30.26	29.91	30.28	25.16	13.63	5.82	2.46
CIFAR100-IDN50	50.3	49.8	50.11	49.91	49.94	45.10	33.64	23.06	15.29
CIFAR100-IDN70	70.05	70.13	70.19	70.04	69.76	65.71	59.47	51.62	44.12
dopanim	39.38	39.54	40.95	40.12	-	38.85	31.82	29.30	-
Flickr_LDL	62.32	58.94	58.10	55.23	50.11	61.09	53.18	47.87	41.99
Chaoyang	13.70	18.68	0.5	-	-	18.01	0.2	-	-

Table 3. Noise ratios(%) of annotators and their majority votes labels on three synthesized and real-world datasets. The № annotator starts with 'a' means the individual annotator in the dataset, the number after 'a' is the serial number of individual annotators, while the 'm' in $\mathrm{m}1 \sim n(n \in 2, 3, 4, 5)$ indicates the noise rates of majority vote labels derived from individual annotator a1, a2, ..., an. For the dopanim and Chaoyang datasets, '-' means that the corresponding items do not exist since the annotators in distinct data versions are different.

Existing multi-rater methods We compare RETINA with the following state-of-the-art/baseline multi-rater methods: 1) **Ensemble** an algorithm that averages the outputs of several single classifiers, each classifier being trained by a regular classifier using the noisy labels from a single annotator; 2) **Majority Vote** which trains one model with a regular classifier using the majority voting labels from all annotators; 3) **CrowdLayer** [36], an end-to-end algorithm that directly trains deep learning models from the noisy labels of multiple annotators using only backpropagation; 4) **FDS** [37] proposed an EM-based algorithm to predict the aggregated consensus labels, then one classifier is trained based on these labels, 5) **Trace-reg** [39] proposed an approach that simultaneously estimates individual annotator reliability through confusion matrices and learns the true label distribution from noisy annotations by incorporating a trace regularization term, 6) **CrowdLab** [17], a two-step algorithm that firstly estimates consensus labels for data examples by aggregating the individual annotations, then a classifier is trained on these consensus labels; 7) **UnionNet** [44], an end-to-end model that maximizes the likelihood of the union of one-hot encoded vectors of labels provided by all annotators with the help of a parametric transition matrix; 8) **Conal** [10], an end-to-end learning solution with two parallel noise adaptation layers that decompose crowdsourced annotation noise into shared confusions across annotators and annotator-specific confusions, 9) **MaDL** [23], an end-to-end algorithm that jointly trains a ground-truth model and an annotator model by presenting a probabilistic training framework; 10) **CrowdAR** [5], an end-to-end algorithm that models the reliability of annotators and is then further used to construct a soft annotation for training; 11) **GeoCrowdNet** [25], an end-to-end system that learns the label correction mechanism and the neural classifer simultaneously; and 12) **Annot-Mix** [24], an algorithm that maximizes the marginal likelihood of observed noisy class labels during the joint training of a classification and an annotator model, thus separating the noise from the true labels.

C. Setting of Experiments

	Backbone	WarmUp Epochs	Epochs	Optimizer	LR Scheduler	Batch Size	Initial_LR
RETINA (DivideMix)	PreAct-ResNet-18	30	300	SGD	Cosine Annealing	128	0.02
RETINA (ProMix)			600			256	0.05
RETINE (Anne)			300			128	0.02

Table 4. Experimental setting of proposed methods on Cifar100-IDN datasets.

	Backbone	WarmUp Epochs	Epochs	Optimizer	LR Scheduler	Batch Size	Initial_LR
RETINA (DivideMix)	Pretrained DINO-V2	1	50	Adam	Cosine Annealing	64	0.02
RETINA (ProMix)			50				
RETINA (Anne)			50				

Table 5. Experimental setting of proposed methods on dopanim dataset.

	Backbone	WarmUp Epochs	Epochs	Optimizer	LR Scheduler	Batch Size	Initial_LR
RETINA (DivideMix)	ResNet-18	10	100	SGD	Cosine Annealing	64	0.02
RETINA (ProMix)			100			128	
RETINA (Anne)			100			64	

Table 6. Experimental setting of proposed methods on Flickr_LDL and Chaoyang.

D. Results

We report the accuracy results of our proposed algorithm, RETINA, on the three synthesized CIFAR100-IDN variants datasets and three real-world datasets, dopanim, Flickr_LDL, and Chaoyang. The experimental results of real-world datasets are as shown below. The bold font indicates the highest accuracy.

	Test Accuracy (%)		
№ annotators	**2**	**3**	**4**
Ensemble	50.74 ± 0.30	51.68 ± 0.28	52.81 ± 0.25
Majority Vote	50.27 ± 0.32	66.98 ± 0.27	77.87 ± 0.31
Crowdlayer [36]	44.89 ± 0.23	59.05 ± 2.14	68.05 ± 3.93
FDS [37]	64.22 ± 0.35	75.93 ± 0.23	76.23 ± 0.21
Trace-reg [39]	46.06 ± 0.55	61.72 ± 0.63	74.35 ± 0.55
CrowdLab [17]	48.89 ± 0.23	65.05 ± 0.64	75.67 ± 0.39
Conal [10]	46.11 ± 0.39	62.03 ± 1.10	75.69 ± 0.14
UnionNet [44]	49.82 ± 0.25	62.77 ± 0.80	74.17 ± 0.19
MaDL [23]	47.76 ± 0.61	64.06 ± 1.33	75.24 ± 1.08
CrowdAR [5]	45.76 ± 0.14	61.51 ± 0.76	73.79 ± 0.31
GeoCrowdNet [25]	50.26 ± 0.27	63.41 ± 0.55	74.31± 0.68
Annot-Mix [24]	51.56 ± 0.13	67.41 ± 0.68	78.30 ± 0.21
RETINA (DivideMix)	51.80 ± 0.21	68.81 ± 0.17	79.31 ± 0.14
RETINA (ProMix)	53.15 ± 0.15	70.27 ± 0.16	80.65 ± 0.15
RETINA (ANNE)	**55.46** ± **0.16**	**71.08** ± **0.13**	**81.77** ± **0.12**

Table 7. Test accuracy (%) on dopanim dataset.

	Test Accuracy (%)			
№ annotators	**2**	**3**	**4**	**5**
Ensemble	40.37 ± 0.55	44.41 ± 0.64	47.17 ± 0.38	48.02 ± 0.40
Majority Vote	37.39 ± 0.53	41.08 ± 0.51	42.22 ± 0.37	46.88 ± 0.34
Crowdlayer [36]	52.46 ± 0.39	53.87 ± 0.55	56.87 ± 0.79	59.77 ±0.90
FDS [37]	42.71 ± 0.51	49.57 ± 0.43	57.22 ± 0.51	58.14 ± 0.46
Trace-reg [39]	47.73 ± 0.43	49.69 ± 0.59	51.39 ± 0.33	53.80 ± 0.36
CrowdLab [17]	44.22 ± 0.76	50.18 ± 0.43	54.76 ± 0.31	58.21 ± 0.20
Conal [10]	48.89 ±0.59	50.11 ±0.22	52.46 ±0.66	54.58 ± 1.02
UnionNet [44]	4.41 ±1.77	6.31 ±0.75	10.52 ±0.83	11.60 ± 2.21
MaDL [23]	47.45 ± 0.28	49.83 ±1.64	52.17 ± 0.82	54.89 ±1.18
CrowdAR [5]	50.12 ± 0.53	50.33 ± 0.91	52.46 ± 0.90	56.26 ± 0.11
GeoCrowdNet [25]	51.23 ±0.46	53.31 ± 0.66	55.31 ±0.31	58.24 ±0.22
Annot-Mix [24]	50.57 ± 1.07	53.02 ± 0.76	55.73 ±0.74	58.76 ± 0.10
RETINA (DivideMix)	57.36 ± 0.44	58.28 ± 0.41	59.12 ± 0.20	60.76 ± 0.11
RETINA (ProMix)	58.64 ± 0.39	59.95 ± 0.26	60.24± 0.14	61.37± 0.10
RETINA (ANNE)	**60.05** ± **0.25**	**61.04**± **0.26**	**61.46** ± **0.18**	**63.45** ± **0.14**

Table 8. Test accuracy (%) on Flickr_LDL dataset.

№ annotators	Test Accuracy (%) 2	3
Ensemble	82.70$\pm$ 0.22	83.22 $\pm$ 0.18
Majority Vote	75.88 $\pm$ 0.31	83.09 $\pm$ 0.17
Crowdlayer [36]	74.16 $\pm$ 0.41	81.39 $\pm$ 0.29
FDS [37]	82.00 $\pm$ 0.20	80.22 $\pm$ 0.19
Trace-reg [39]	80.69 $\pm$ 0.27	84.32 $\pm$ 0.16
CrowdLab [17]	81.45 $\pm$ 0.24	83.26 $\pm$ 0.18
Conal [10]	81.49 $\pm$ 0.28	75.33 $\pm$ 0.35
UnionNet [44]	80.27 $\pm$ 0.30	81.81 $\pm$ 0.22
MaDL [23]	80.78 $\pm$ 0.23	83.37 $\pm$ 0.17
CrowdAR [5]	79.15 $\pm$ 0.32	81.76 $\pm$ 0.20
GeoCrowdNet [25]	78.82 $\pm$ 0.36	83.77 $\pm$ 0.29
Annot-Mix [24]	79.15 $\pm$ 0.18	81.90 $\pm$ 0.13
RETINA (DivideMix)	83.18 $\pm$ 0.13	84.89 $\pm$ 0.09
RETINA (ProMix)	83.59 $\pm$ 0.11	85.09 $\pm$ 0.08
RETINA (ANNE)	**83.42** $\pm$ **0.10**	**85.61** $\pm$ **0.07**

Table 9. Test accuracy (%) on Chaoyang dataset.

E. Ablation Study

№ annotators	**2**	**3**	**4**	**5**
CIFAR100-IDN70				
Biased Selection	72.65 $\pm$ 0.64	74.56 $\pm$ 0.41	**75.38** $\pm$ **0.25**	**75.69** $\pm$ **0.24**
Random Selection	**73.00** $\pm$ **0.65**	**74.63** $\pm$ **0.46**	75.03 $\pm$ 0.44	75.41 $\pm$ 0.37
dopanim				
Biased Selection	51.32 $\pm$ 0.27	67.62 $\pm$ 0.46	**79.32** $\pm$ **0.22**	-
Random Selection	**51.80** $\pm$ **0.21**	**68.81** $\pm$ **0.17**	79.31 $\pm$ 0.10	-

Table 10. Test accuracy (%) on different teacher-student selection algorithms based on: (top) CIFAR100-IDN70, and (bottom) dopanim.