SUPPLEMENTARY OF LEARNING WITH INSTANCE-DEPENDENT NOISY LABELS BY ANCHOR HALLUCINATION AND HARD SAMPLE LABEL CORRECTION

Po-Hsuan Huang[1,⋆] *Chia-Ching Lin*[1,⋆] *Chih-Fan Hsu*[1] *Ming-Ching Chang*[2] *Wei-Chao Chen*[1]

[1]Inventec Corporation, Taipei, Taiwan
[2]Department of Computer Science, University at Albany, State University of New York, NY, USA

A. IMPLEMENTATION DETAILS

Table 1 shows the hyperparameter settings for different datasets in our experiment. We train our models separately on (1) CIFAR-10 dataset with synthetic Instance Dependent Noise (IDN), (2) the CIFAR-10N, (3) CIFAR-100N, and (4) Clothing1M datasets. The training process of our model spanned 200 epochs using SGD with an initial learning rate of 0.02, a momentum of 0.9, a weight decay parameter of 0.0005, and a batch size of 128. The number of warm-up epochs is set as 10 for CIFAR-10 and 30 for CIFAR-100. At epoch 120, we divide the learning rate by 10. The $P\%$ portion of easy sample selection is set to 0.4, 0.7, and 0.8 for Classification-based label noise with 10% and 40% noise ratios and CIFAR-100N, respectively. For the PTD label noise with both 20% and 40% noise ratios and for CIFAR-10N, we set $P\%$ as 0.6. The hard sample selection threshold λ_{conf} is set as 0.95 for CIFAR-10N, and 0.8 for both 40% Classification-based label noise and CIFAR100-N. For 10% Classification-based label noise and all the noise ratios in PTD, λ_{conf} is set as 0.97. As for λ_{mse}, we simply follow the value suggested in DivideMix [4]. We evaluate our method on a clean testing set and report the best testing accuracy on the average of three different trials.

For Clothing1M, we train the model for 80 epochs using SGD with an initial learning rate of 0.02, a momentum of 0.9, a weight decay parameter of 0.001, and a batch size of 32. The number of warm-up epochs is set as 1. At epoch 40, we divide the learning rate by 10. The threshold of GMM-based easy sample selection is set to 0.5, and the hard sample selection threshold λ_{conf} is set as 0.9. We again follow DivideMix [4] and set $\lambda_{conf} = 0$. During training, we use the 14K clean validation set to choose the best model, which is applied to the 10K clean test to get the test accuracy.

Throughout all experiments, the difficulty level λ_p is fixed at 0.6, and the maximum number of valid representatives for each real hard sample (K) is set as 3.

⋆Equal contribution

B. HYPERPARAMETER ANALYSIS

We conduct additional experiments on CIFAR-10 with 40% classification-based noise to examine the effect of the two hyperparameters: the threshold for easy sample selection ($P\%$ in § 3.1) and the threshold for hard sample selection (λ_{conf} in § 3.3).

Threshold for easy sample selection. In step 1 (§ 3.1), we select a fixed portion of samples with top-$P\%$ easiness scores to form the easy feature set $\mathcal{S}_e$ for the subsequent hard anchor hallucination and hard sample selection processes. Intuitively, with a larger P, $\mathcal{S}_e$ would have sufficient samples for each class, but might include more noisy samples. In Table 2, we show the test accuracy of the model trained on CIFAR-10 with 40% classification-based noise, with $P\% \in \{0.3, 0.4, 0.5\}$ and a fixed $\lambda_{conf} = 0.8$. We can observe that the model's performance deteriorates when $P\% = 0.3$, as the easy set $\mathcal{S}_e$ might not contain sufficient samples for all classes. On the other hand, when $P\%$ surpasses a certain threshold (e.g., 0.4), the model consistently achieves high performance and shows less sensitivity to the size of $\mathcal{S}_e$, indicating the robustness of the proposed framework.

Threshold for hard sample selection. In the step of hard sample selection(§3.3), we define *clean hard* samples from the hard feature set $\mathcal{S}_h$ based on the cosine similarity values between real hard features and the hallucinated anchors. Specifically, an hallucinated anchor s_a is defined as a *valid* representative of a real hard feature s_r if $\langle s_a, s_r\rangle \geq \lambda_{conf}$. Intuitively, a smaller λ_{conf} would result in a larger size of selected clean hard set $\mathcal{S}_h^c$, but might introduce more noisy hard samples. In Table 3, we show the test accuracy of the model trained on CIFAR-10 with 40% classification-based noise, with $\lambda_{conf} \in \{0.7, 0.8, 0.9\}$ and a fixed $P\% = 0.4$. We observe that the model performance exhibits notable variations based on the selection of different values for λ_{conf}. This implies that both the quantity and quality of the selected clean hard samples $\mathcal{S}_h^c$ are crucial for the model performance and the precise tuning of λ_{conf} is necessary.

Table 1. Hyperparameter settings.

Dataset	CIFAR-10				CIFAR-10N		CIFAR-100N	Clothing1M
Noise type	C-based 10/20%	C-based 40%	PTD-20%	PTD-40%	Random 1/2/3	Worst	Noisy	
Total epochs	200	200	200	200	300	200	200	80
Warm-up epochs	10	10	10	10	10	10	30	1
Init. learning rate	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.02
SGD Momentum	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Weight decay	5e-4	5e-4	5e-4	5e-4	5e-4	5e-4	5e-4	1e-3
Batch size	128	128	128	128	128	128	128	32
λ_p	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
K	3	3	3	3	3	3	3	3
$P\%$	0.4	0.7	0.6	0.6	0.8	0.6	0.8	0.7
λ_{conf}	0.97	0.8	0.97	0.97	0.95	0.95	0.8	0.9
λ_{mse}	0	25	25	25	0	25	150	0

Table 2. Hyperparameter analysis of the fixed proportion of easy sample selection $P\%$ on CIFAR-10 with 40% Classification-based noise and $\lambda_{conf} = 0.8$.

$P\%$	0.3	0.4	0.5
Test Acc.	90.38±0.51	**92.47**±0.41	92.38±0.14

Table 3. Hyperparameter analysis of the hard sample selection threshold λ_{conf} on CIFAR-10 with 40% Classification-based noise and $P\% = 0.4$.

λ_{conf}	0.7	0.8	0.9
Test Acc.	91.12±0.76	**92.47**±0.41	91.53±1.18

C. VISUALIZATION

Visualization for hard sample selection. Our hallucinator generates hard sample anchors in a feature space that is not intuitive to observe. To demonstrate the efficacy of such hard anchor hallucination and sample selection, we search for the nearest real samples in the feature space and take them as visual substitutes. Fig. 1 shows such visualization results on CIFAR-10, where each combination of the input easy sample pairs and their hallucination anchors are shown. Observe that our hallucinator can effectively identify challenging samples with correct labels (as shown in the first column) and rectify samples with incorrect labels (as evident in the fifth column). This experiment provides additional evidence of the ability of our hallucinator to produce high-quality anchors and reinforces the practical utility of our method.

Performance of noise correction. We show the overall noise rate and the label correction accuracy of our method on the most challenging CIFAR-10 with 40% Classification-based noise during the training in Fig. 2. The overall noise ratio decreased during training and our overall label correction steadily achieved over 90% correction accuracy, which shows the effectiveness of our method.

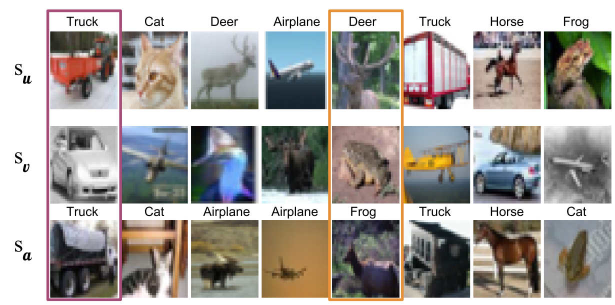

Fig. 1. **Visual verification of the hard anchor selection process.** The first two rows represent the corresponding images for easy features s_u and s_v sampled from $\mathcal{S}_e$, and the third row represents the nearest image to the hallucinated anchor $s_a = h_\phi(s_u, s_v)$. The first column (violet box) shows that s_a successfully selects the correctly labeled real hard sample (*Truck*). The fifth column (orange box) shows that s_a successfully corrects the label of an incorrectly labeled real hard sample (*Frog* to *Deer*).

Examples of hard samples in CIFAR-10. We present some of the hard examples from CIFAR-10 in Fig. 3. These samples exhibit notable difficulty as they often bear resemblance to other classes or with hard visual patterns. For example, the background of the first sample could potentially lead to a misclassification of the ship.

D. PSEUDO CODE FOR OUR MODEL TRAINING PROCEDURE

We provide the pseudo-code for our framework in Algorithm 1 for model training.

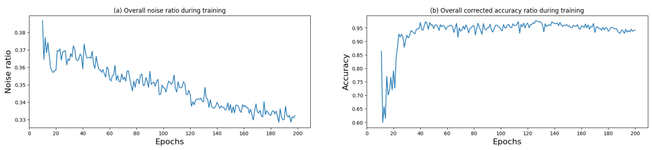

Fig. 2. The correction performance and noise curves. The left figure (a) shows the overall noise rate gradually decreased during training. The right figure (b) is our overall correction accuracy.

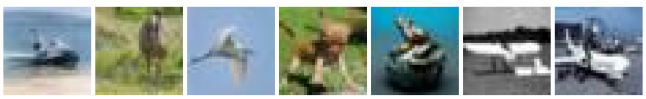

Fig. 3. Some hard examples in CIFAR-10. These samples are easily being confused with other classes because of their hard visual patterns.

Algorithm 1 The proposed training procedure.

Input: The training set $\mathcal{D} = \{(x_n, \tilde{y}_n)\}$, number of class C, classification network $f_\theta \circ g_\rho$, hallucinator h_ϕ, easy selection threshold $P\%$, hard sample selection threshold λ_{conf}, total training epochs T, number of iterations I_{max}, number of warm-up epochs T_{warm}, learning rate η

Output: Trained model $f_\theta \circ g_\rho$

for $t = 1, 2, \ldots, T$ **do**
  if $t \leq T_{warm}$ **then**
    Update $(\theta, \rho) \leftarrow (\theta, \rho) - \eta \nabla \mathcal{L}_{CE}$
  else
    Freeze (θ, ρ) and un-freeze ϕ
    Get easiness score ω_n for all samples $\in \mathcal{D}$ from GMM
    Get sets of $\mathcal{S}_e$ and $\mathcal{S}_h$ by Eq. (1)
    Initialize $\mathcal{S}_{hal}$ as an empty set
    for $iter = 1, 2, \ldots, I_{max}$ **do**
      Sample a mini-batch S from $\mathcal{S}_e$
      Hallucinate anchors $\{s_a\}$ from S
      $\mathcal{S}_{hal} \leftarrow \mathcal{S}_{hal} \cup \{s_a\}$
      Obtain $\mathcal{L}_{hal}$ using $\{s_a\}$ and S by Eq. (2)
      Update $\phi \leftarrow \phi - \eta \nabla \mathcal{L}_{hal}$
    end for
    Freeze ϕ and Un-freeze (θ, ρ)
    Select $\mathcal{S}_h^c$ from $\mathcal{S}_h$ using $\mathcal{S}_{hal}$
    $\mathcal{S}_{labeled} \leftarrow \mathcal{S}_e \cup \mathcal{S}_h^c$
    $\mathcal{S}_{unlabeled} \leftarrow \mathcal{S}_h \setminus \mathcal{S}_h^c$
    for $iter = 1, 2, \ldots, I_{max}$ **do**
      Obtain $\mathcal{L}_{CE}$ using $\mathcal{S}_{labeled}$ and $\mathcal{S}_{unlabeled}$ by Eq. (3)
      Update $(\theta, \rho) \leftarrow (\theta, \rho) - \eta \nabla \mathcal{L}_{CE}$
    end for
  end if
end for