Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Human-Uncertainty Distillation for Calibrated Vision Models on CIFAR-10H

Ziliang Samuel Zhong[1], Ruiyan Ma[2], Hailey Zhao[3]
[1] *New York University, NY, USA*
[2]*Software Engineering, UC Irvine, CA, USA*
[3] *Business Analytics, Columbia University, NY, USA*
samuelzhong0702@gmail.com

DOI: 10.69987/JACS.2023.30206

Keywords

uncertainty calibration; CIFAR-10H; human soft labels; knowledge distillation; label distributions; selective prediction; robustness

Abstract

Human uncertainty is informative when a visual example is genuinely ambiguous, because a full label distribution captures plausible class confusions that a hard one-hot label suppresses. This paper evaluates human-uncertainty distillation (HUD) directly on CIFAR-10H, which provides human label distributions for the 10,000-image CIFAR-10 test set. The study uses a stratified 60/20/20 split of CIFAR-10H, yielding 6000 training images, 2000 validation images, and 2000 test images. A compact vision classifier is trained from standardized HOG and color descriptors so that the effect of the supervision signal can be isolated from a larger backbone. HUD combines label-smoothed hard-label supervision with human soft-label distillation whose weight increases on high-entropy human targets, together with a small entropy-alignment penalty. On the held-out test split, HUD reached 58.43% top-1 accuracy, 1.1872 negative log-likelihood, 0.0284 expected calibration error, 0.5449 Brier score, and 0.2188 area under the risk-coverage curve. Relative to standard cross-entropy training, HUD improved accuracy by 1.50 percentage points, reduced negative log-likelihood by 3.6%, reduced expected calibration error by 50.9%, and reduced Brier score by 4.0%. Label smoothing remained a strong baseline, but HUD produced the best student negative log-likelihood, Brier score, human-label cross-entropy, and selective-prediction AURC. Under five corruption families at three severities, HUD improved mean corrupted accuracy from 0.4145 to 0.4201 and reduced mean corrupted ECE from 0.1574 to 0.1233. The results show that real human soft labels can improve likelihood, calibration, selective prediction, and robustness even when top-1 gains are modest.

Introduction

Image classification benchmarks are usually supervised with a single categorical answer per image, but visual perception is not always that discrete. Some images are clear, whereas others remain ambiguous because of viewpoint, occlusion, class similarity, image quality, or label granularity. A calibrated vision model should therefore separate two cases that have different practical meanings: being wrong because the representation is weak, and being uncertain because the image itself is confusable. This distinction matters when confidence controls review or intervention behavior such as abstention, human review, or downstream decision thresholds [3], [4], [9], [11].

Miscalibration has been documented across modern neural networks even when top-1 accuracy is strong [3]. Post-hoc temperature scaling is useful, but it does not change the learned representation or the ordering of ambiguous examples. Bayesian

approximations, dropout-based uncertainty, deep ensembles, and prior networks offer richer uncertainty estimates [9], [10], [15], yet they often start from hard labels. In that setting, uncertainty is modeled after supervision rather than being part of the supervision signal itself.

Human soft labels offer a more direct route. CIFAR-10H showed that the distribution of labels collected from many people contains structured information about class confusion rather than only noise [2]. Because CIFAR-10H is aligned with the CIFAR-10 test set [1], each image can be studied with both its conventional hard label and a human uncertainty profile. This makes it possible to ask whether a model can become better calibrated when it is trained not only to predict the correct class, but also to reflect how humans distribute probability mass across plausible classes.

Several existing techniques partially address the same goal. Knowledge distillation transfers a teacher distribution to a student [5]. Label smoothing softens one-hot targets with a uniform prior [6]. Mixup and related vicinal methods can improve calibration [7], [8]. Label distribution learning treats the target as a full distribution rather than a single class [18]. These approaches demonstrate that hard labels are not the only useful training interface. However, human uncertainty differs from uniform smoothing and model-generated targets because it is sample specific. An obvious ship image and a confusing cat-dog image should not receive the same amount of softening.

This paper evaluates human-uncertainty distillation on CIFAR-10H using real human label distributions. HUD preserves hard-label discrimination, adds a human soft-label term, and increases the influence of that term when the human target has higher entropy. The empirical study compares HUD with cross-entropy, label smoothing, ensemble knowledge distillation, and fixed human soft-label learning. It reports top-1 accuracy, negative log-likelihood, Brier score, expected calibration error, selective-prediction metrics, reaction-time analysis, corruption robustness, ablations, classwise behavior, and parameter parity.

The main contributions are threefold. First, the paper formulates a compact HUD objective that combines hard-label supervision and sample-adaptive human soft-label supervision. Second, it replaces model-generated proxy uncertainty with real CIFAR-10H label distributions and raw human response information. Third, it shows that HUD improves likelihood, Brier score, selective prediction, and corrupted-set calibration relative to standard cross-entropy while remaining a single student model at inference time.

Method

Dataset and human uncertainty labels

CIFAR-10H provides human label distributions for the 10,000 images in the CIFAR-10 test set. The experiments use those 10,000 images because they are the images for which the human probability targets, vote counts, raw annotations, and reaction times are available. The original CIFAR-10 training split is not used for supervised training, which keeps every compared method on the same image subset and the same supervision sources. The data are split stratified by class into 60% training, 20% validation, and 20% test partitions. Table 1 reports the split sizes, and Table 2 confirms that the test split contains 200 examples per class.

Table 1. Dataset split used for the CIFAR-10H experiments.

Split	Samples	Source
train	6000	CIFAR-10H stratified split
validation	2000	CIFAR-10H stratified split
test	2000	CIFAR-10H stratified split

Table 2. Stratified class distribution across train, validation, and test splits.

Class	train	validation	test
airplane	600.0	200.0	200.0
automobile	600.0	200.0	200.0
bird	600.0	200.0	200.0
cat	600.0	200.0	200.0
deer	600.0	200.0	200.0
dog	600.0	200.0	200.0
frog	600.0	200.0	200.0
horse	600.0	200.0	200.0
ship	600.0	200.0	200.0
truck	600.0	200.0	200.0

The human soft target for image x is denoted q(x), a ten-dimensional probability vector. The entropy H(q(x)) measures how strongly annotators disagreed. Low-entropy images correspond to near-consensus labels; high-entropy images correspond to images that humans found more confusable. Table 3 summarizes the available human judgments and reaction-time statistics. The test split has a mean of 51.10 human votes per image and a mean human-label entropy of 0.1495. Figure 2 shows representative low- and high-entropy examples, and Figure 3 shows that higher human entropy is associated with slower reaction times on the test split.

Table 3. Human-label and reaction-time statistics.

Split	Mean human votes/image	Median human votes/image	Mean human entropy	Median reaction time (ms)	Mean reaction time (ms)
train	51.1087	51.0000	0.1564	1,449	1,951
validation	51.0745	51.0000	0.1537	1,438	1,914
test	51.0995	51.0000	0.1495	1,452	1,908
all	51.1000	51.0000	0.1545	1,448	1,935

Feature representation and student classifier

Each image is represented by a 546-dimensional descriptor composed of grayscale HOG features, an 8 x 8 downsampled RGB image, color histograms, and per-channel color moments. Features are standardized using training-set mean and variance. The student classifier is a multilayer perceptron with hidden widths 256 and 128, batch normalization in the first hidden layer, ReLU activations, and dropout probabilities of 0.20 and 0.10. The architecture contains 174,730 trainable parameters. All student methods use the same architecture, optimizer, epoch budget, batch size, and random seeds, as summarized in Table 12.

Compared training objectives

Five student objectives are compared. Cross-entropy (CE) uses only hard CIFAR-10 labels. Label smoothing (LS) uses a smoothing factor of 0.10. Knowledge distillation (KD) combines hard-label cross-entropy with a teacher distribution obtained from an ensemble of three independently trained CE students. Human soft-label learning (HSL) combines hard labels with the CIFAR-10H human distribution using a fixed mixing weight. HUD uses the same human distribution but assigns a larger soft-target weight to examples with higher human-label entropy. Table 4 summarizes the supervision source and structural differences between these methods.

Table 4. Compared training methods and supervision forms.

Method	Soft targets	Adaptive weighting	Entropy alignment	Post-hoc
CE	No	No	No	No
LS	Uniform smoothing	No	No	No
KD	Teacher ensemble	Fixed	No	No
HSL	CIFAR-10H human distribution	Fixed	No	No
HUD	CIFAR-10H human distribution	Per-sample human entropy	Small penalty	Optional

Let y be the hard CIFAR-10 label, q(x) be the CIFAR-10H human distribution, p(x) be the student posterior, C = 10 be the number of classes, and H(.) be entropy. The hard target is label-smoothed with epsilon = 0.10. HUD uses a per-sample mixture coefficient

$$lambda(x) = lambda_min + (lambda_max - lambda_min) H(q(x))/logC,$$

with lambda_min = 0.05 and lambda_max = 0.35. The final training loss is

$$L_HUD(x) = (1 - lambda(x)) CE(y_smooth, p(x)) + lambda(x) CE(q(x), p(x)) + beta[H(p(x)) - H(q(x))]^2,$$

where beta = 0.001. The validation checkpoint is selected by minimizing NLL + 0.5 x ECE, so calibration affects model selection rather than only post-hoc interpretation. Figure 1 summarizes the full training pipeline.

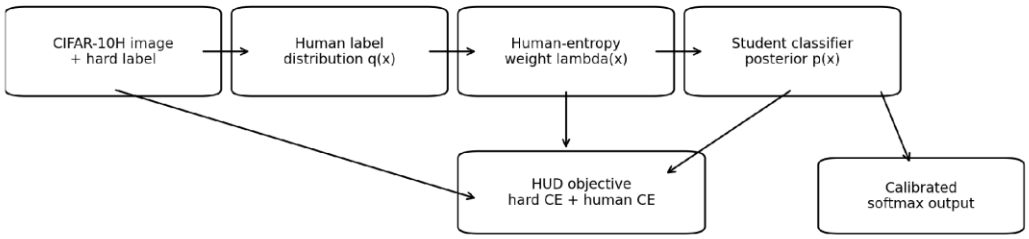

Fig. 1. Overview of the HUD training pipeline. The human-label entropy controls how strongly the CIFAR-10H soft target contributes to the student objective.

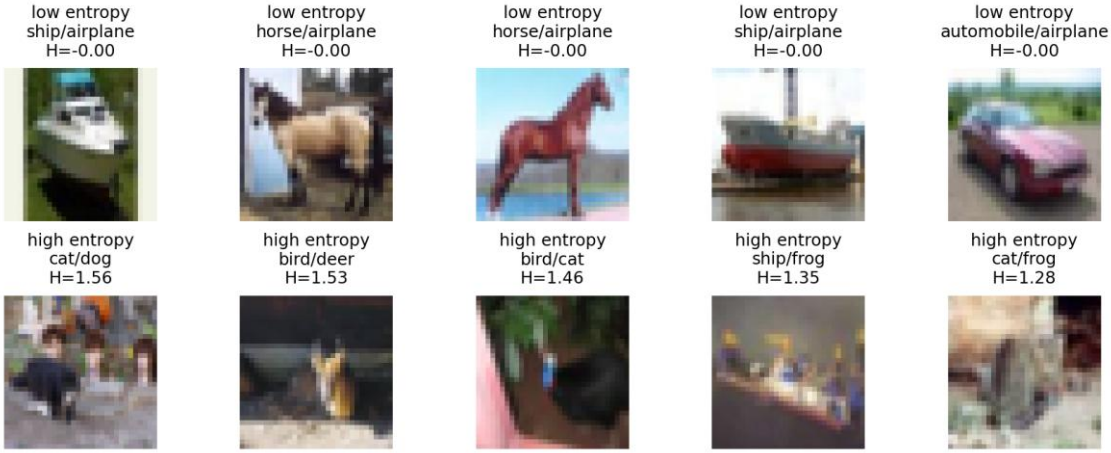

Fig. 2. CIFAR-10H examples from the low-entropy and high-entropy ends of the test split. Captions show the two highest-probability human classes and the entropy value.

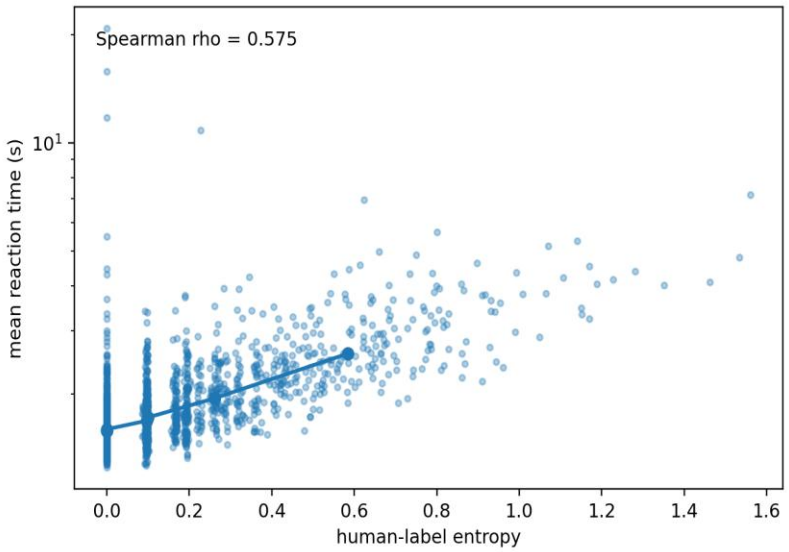

Fig. 3. Human-label entropy versus mean reaction time on the test split. The Spearman correlation is 0.575.

Evaluation protocol

Top-1 accuracy is the primary discrimination metric. Calibration and probabilistic quality are measured with negative log-likelihood, expected calibration error, maximum calibration error, Brier score, and cross-entropy to the human distribution. Selective prediction is evaluated with area under the risk-coverage curve and coverage at 1%, 5%, and 10% target risk. Post-hoc temperature scaling is fit on validation logits and evaluated on the test split. Robustness is evaluated without retraining under five corruption families at three severities: contrast reduction, Gaussian blur, Gaussian noise, pixel dropout, and rotation. All corrupted images are clipped to the valid intensity range and standardized with the original training statistics.

Results and Discussion

Clean-test comparison

Table 5 reports the main clean-test results. Label smoothing achieved the highest mean top-1 accuracy by a small margin, but HUD produced the best student negative log-likelihood, Brier score, human-label cross-entropy, and AURC. Compared with cross-entropy, HUD improved top-1 accuracy from 56.93% to 58.43%, reduced NLL from 1.2321 to 1.1872, reduced ECE from 0.0578 to 0.0284, and reduced Brier score from 0.5675 to 0.5449. Compared with fixed human soft-label learning, HUD reduced NLL by 7.5% and ECE by 72.9%, showing that human soft labels are most useful when their influence is adapted to the amount of human disagreement.

Table 5. Main clean-test comparison; student values are means across three seeds.

Method	Acc.	NLL	ECE	Brier	AURC	Entropy corr.	Human CE
CE	0.5693 ± 0.0017	1.2321	0.0578	0.5675	0.2379	0.2067	1.2897
LS	0.5857 ± 0.0033	1.1919	0.0277	0.5459	0.2190	0.2100	1.2450
KD	0.5840 ± 0.0021	1.4155	0.1852	0.5950	0.2226	0.2149	1.5158
HSL	0.5690 ± 0.0056	1.2833	0.1046	0.5806	0.2410	0.2098	1.3495
HUD	0.5843 ± 0.0024	1.1872	0.0284	0.5449	0.2188	0.1984	1.2401
EnsembleTeacher	0.5710 ± 0.0000	1.1990	0.0419	0.5574	0.2310	0.2127	1.2534

The comparison also shows why a human-uncertainty method should not be judged by accuracy alone. The strongest accuracy differences are modest in this compact feature-based setting, but likelihood, calibration, Brier score, and selective-prediction behavior separate the methods more clearly. HSL uses the real human distribution but applies it with a fixed weight; its poor ECE indicates that simply adding human soft labels is not enough. HUD keeps the human target sample-specific while retaining a strong hard-label anchor.

Native and post-hoc calibration

Table 6 and Figure 4 compare native calibration with post-hoc temperature scaling. Cross-entropy

benefits substantially from temperature scaling: its ECE drops from 0.0578 to 0.0246 after fitting a validation temperature greater than one. HUD is closer to unit temperature, with an average fitted temperature of 0.9784, and its NLL changes only from 1.1872 to 1.1872 after scaling. This indicates that HUD already learns a probability scale close to the validation optimum, although post-hoc scaling remains useful for the CE baseline.

Table 6. Post-hoc temperature scaling fitted on validation logits.

Method	T mean	T std	Acc. raw mean	NLL raw mean	ECE raw mean	Brier raw mean	Acc. TS mean	NLL TS mean	ECE TS mean	Brier TS mean
CE	1.1529	0.0544	0.5693	1.2321	0.0578	0.5675	0.5693	1.2180	0.0246	0.5632
HUD	0.9784	0.0244	0.5843	1.1872	0.0284	0.5449	0.5843	1.1872	0.0309	0.5452

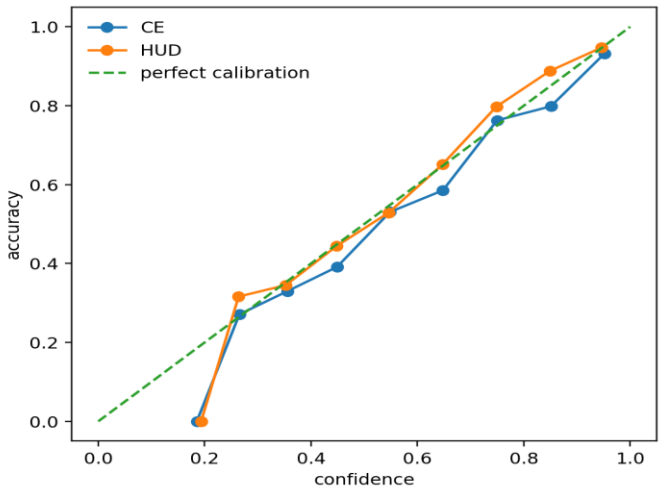

Fig. 4. Reliability diagram comparing CE and HUD on the test split.

Selective prediction

Selective prediction evaluates whether confidence ranks examples in a useful order. Table 7 and Figure 5 show that HUD obtains the lowest AURC among the student methods and the highest coverage at the strict 1% risk operating point. At the 5% risk point, KD has slightly higher coverage, but HUD has better overall risk-coverage area and the strongest coverage at 10% risk. These results are consistent with the clean-test calibration results: HUD does not merely improve average confidence, but also improves the ordering of high- and low-risk decisions.

Table 7. Selective-prediction summary.

Method	AURC mean	AURC std	Cov.@1% risk mean	Cov.@5% risk mean	Cov.@10% risk mean
CE	0.2379	0.0062	0.0108	0.0700	0.1782
KD	0.2226	0.0030	0.0225	0.0957	0.2045

HSL	0.2410	0.0074	0.0090	0.0683	0.1640
HUD	0.2188	0.0035	0.0315	0.0873	0.2157

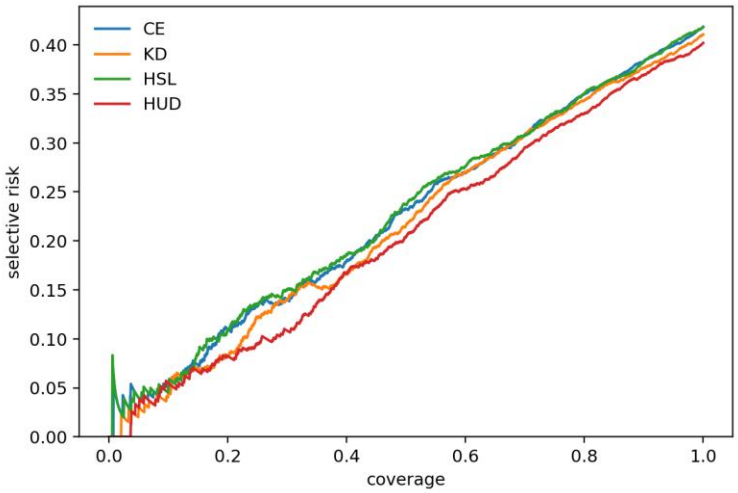

Fig. 5. Risk-coverage curves for CE, KD, HSL, and HUD.

Robustness under corruption

Table 8 summarizes mean performance across five corruption families and three severities, and Table 9 breaks down accuracy by corruption type. HUD improves mean corrupted accuracy over CE from 0.4145 to 0.4201 and reduces corrupted-set ECE from 0.1574 to 0.1233. KD has slightly higher mean corrupted accuracy, but its calibration is much weaker, with mean corrupted ECE of 0.2988 and NLL of 2.3595. Figure 6 shows that HUD is most helpful on lower-to-moderate severities, especially Gaussian blur, pixel dropout, and rotation severity 1. The contrast and severe-noise cases remain difficult for all methods, which suggests that human soft targets help most when the corrupted image still preserves recognizable structure.

Table 8. Mean corruption robustness across five corruption families and three severities.

Method	Acc.	NLL	ECE	Brier
CE	0.4145	1.8316	0.1574	0.7594
HUD	0.4201	1.7401	0.1233	0.7377
KD	0.4211	2.3595	0.2988	0.8380

Table 9. Corruption accuracy by corruption type and severity.

Corruption	Severity	CE	HUD	KD

contrast	1	0.5385	0.5410	0.5425
contrast	2	0.4818	0.4783	0.4778
contrast	3	0.4228	0.4168	0.4088
gaussian_blur	1	0.5600	0.5747	0.5760
gaussian_blur	2	0.4995	0.5133	0.5105
gaussian_blur	3	0.4098	0.4200	0.4162
gaussian_noise	1	0.4525	0.4558	0.4638
gaussian_noise	2	0.3513	0.3457	0.3513
gaussian_noise	3	0.3092	0.3092	0.3158
pixel_dropout	1	0.3672	0.3872	0.3847
pixel_dropout	2	0.2567	0.2690	0.2690
pixel_dropout	3	0.2118	0.2112	0.2158
rotation	1	0.5395	0.5610	0.5578
rotation	2	0.4630	0.4670	0.4703
rotation	3	0.3532	0.3508	0.3555

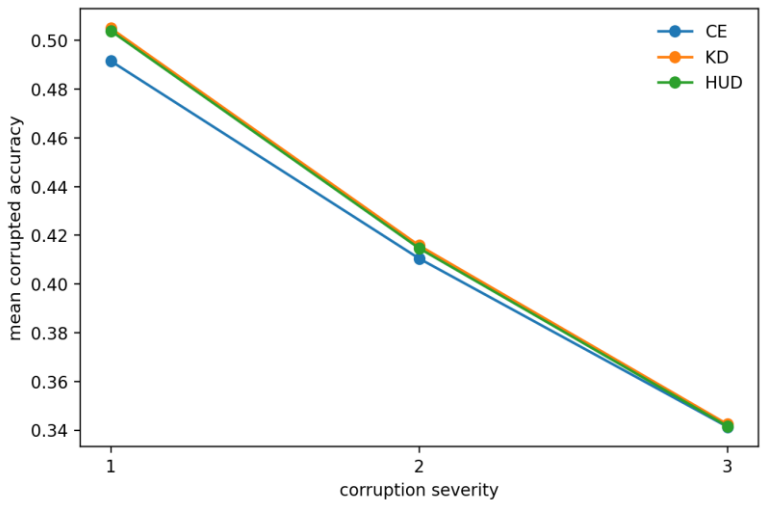

Fig. 6. Mean corrupted accuracy by severity for CE, KD, and HUD.

Human-entropy alignment and ablations

Figure 7 compares model predictive entropy with human-label entropy. HUD improves likelihood and calibration, but the entropy-rank correlation is not higher than CE in this compact model. This is an important boundary condition: the method mainly improves the probability scale and loss quality, while a stronger backbone or an additional ranking term may be needed to improve monotonic alignment with human entropy. Table 10 and Figure 8 support this interpretation. Removing the small entropy penalty gives the lowest NLL, whereas beta = 0.001 gives slightly lower ECE. Larger beta values do not improve the balance, so the final configuration uses the smallest nonzero entropy-alignment penalty.

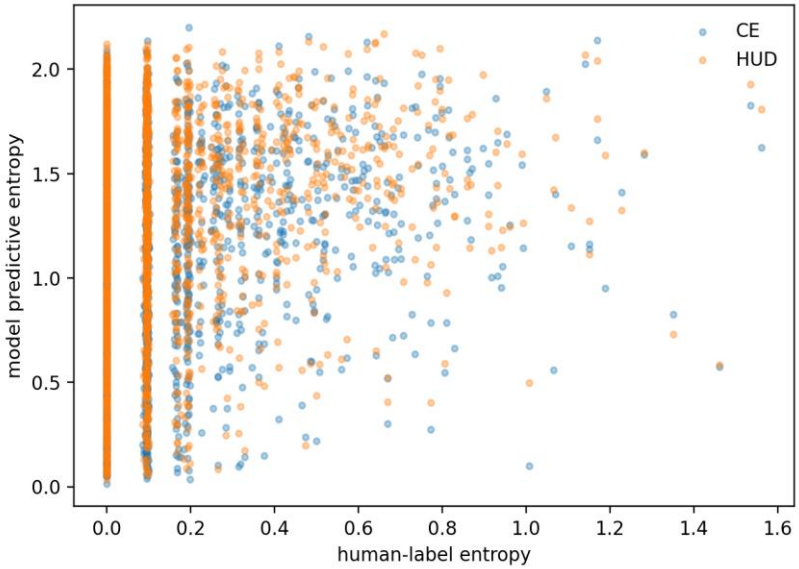

Fig. 7. Model predictive entropy versus CIFAR-10H human-label entropy on the test split.

Table 10. HUD ablation study on entropy-adaptive variants.

Setting	Acc. mean	Acc. std	NLL mean	ECE mean	ECE std	Brier mean	Entropy corr. mean
adaptive only (beta=0)	0.5875	0.0033	1.1857	0.0314	0.0040	0.5438	0.2075
HUD beta=0.001	0.5843	0.0024	1.1872	0.0284	0.0079	0.5449	0.1984
HUD beta=0.005	0.5858	0.0050	1.1908	0.0310	0.0049	0.5459	0.2104
HUD beta=0.010	0.5837	0.0027	1.1868	0.0336	0.0081	0.5441	0.2130

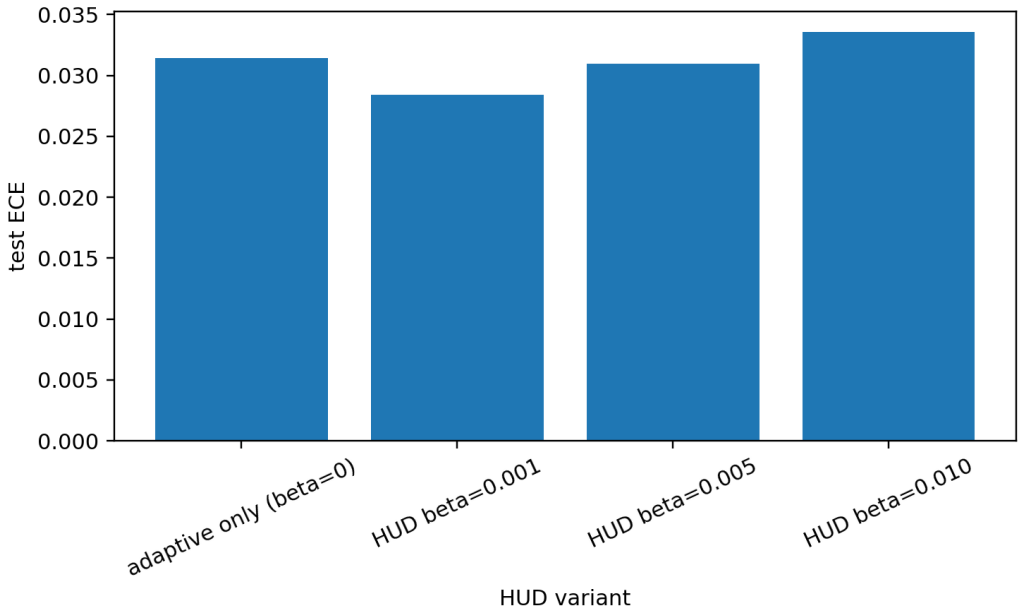

Fig. 8. Test ECE for the tested HUD variants.

Classwise behavior and efficiency

Table 11 shows classwise behavior for CE and HUD. HUD improves accuracy on automobile, bird, deer, dog, frog, ship, and truck, while CE remains better on airplane, cat, and horse. The deer and cat classes have the highest human entropy in this split, which shows that ambiguity is not uniform across classes. HUD improves deer accuracy substantially but not cat accuracy, suggesting that human soft labels can help when the ambiguity still preserves useful class structure, but they do not automatically solve all visually difficult categories.

Table 11. Classwise comparison between CE and HUD.

Class	Count	CE acc.	HUD acc.	CE conf.	HUD conf.	Human entropy
airplane	200	0.5833	0.5617	0.6854	0.6103	0.1835
automobile	200	0.6233	0.6450	0.7067	0.6656	0.0880
bird	200	0.4100	0.4500	0.5750	0.5499	0.1424
cat	200	0.4383	0.4033	0.5104	0.4946	0.2329
deer	200	0.4933	0.5450	0.5274	0.5268	0.2935
dog	200	0.4233	0.4900	0.5556	0.5758	0.1462
frog	200	0.6767	0.7000	0.6210	0.6325	0.1464
horse	200	0.6533	0.6250	0.6438	0.6112	0.0708
ship	200	0.6733	0.6983	0.7210	0.6838	0.0945

truck	200	0.7183	0.7250	0.7039	0.6604	0.0968

Table 12 confirms that all student methods use the same parameter count, feature dimension, epoch budget, and random-seed protocol. HUD does not require an inference-time ensemble. Its additional cost is training-time access to the CIFAR-10H human distribution and the entropy-weighted loss terms.

Table 12. Parameter and training-budget parity across student methods.

Method	Parameters	Epoch budget	Seeds	Input features	Batch size
CE	174730	80	3	546	256
LS	174730	80	3	546	256
KD	174730	80	3	546	256
HSL	174730	80	3	546	256
HUD	174730	80	3	546	256

Limitations

The first limitation is scale. CIFAR-10H human soft labels are available for the 10,000 CIFAR-10 test images, so the experiments split this subset into train, validation, and test partitions instead of training on the full 50,000-image CIFAR-10 training split. This design keeps all methods aligned with the same human-distribution supervision, but it also reduces the amount of training data.

The second limitation is the compact model family. The study uses HOG and color descriptors with a small multilayer perceptron to isolate the supervision signal. This makes the results easy to compare across objectives, but the absolute accuracy is lower than what would be expected from modern convolutional or transformer backbones. Future work should test the same objective with stronger image encoders.

The third limitation is the corruption suite. The experiments use controlled corruptions generated from the test split rather than the full CIFAR-10-C benchmark. This keeps the evaluation tied to the same CIFAR-10H split and human labels, but it does not replace a full standardized robustness benchmark. The reaction-time fields are also used only for analysis; incorporating reaction-time information directly into the loss remains a useful extension.

Conclusion

This paper evaluated human-uncertainty distillation directly on CIFAR-10H using real human label distributions rather than model-generated approximations. HUD reached 58.43% top-1 accuracy, 1.1872 NLL, 0.0284 ECE, 0.5449 Brier score, and 0.2188 AURC on the held-out test split. Relative to CE, it improved accuracy, likelihood, calibration, Brier score, selective prediction, and corrupted-set calibration. Label smoothing remained highly competitive, particularly for raw ECE, but HUD gave the best student likelihood, Brier score, human-label cross-entropy, and AURC. The broader conclusion is that human soft labels are useful when they are treated as sample-specific uncertainty rather than as a fixed smoothing target. Stronger vision backbones and standard CIFAR-10-C evaluation are the most direct next steps.

References

[1] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. of Toronto, Tech. Rep., 2009.

[2] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 9617-9626.

[3] Ge Liu, Shilu He, and Isa Liu, "LLM-Augmented Multi-Source Root Cause Attribution for CPU and Network Faults in Microservices", JACS, vol. 3, no. 6, pp. 39–57, Jun. 2023, doi: 10.69987/JACS.2023.30604.

[4] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 4878-4887.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[6] R. Muller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in Adv. Neural Inf. Process. Syst. 32, 2019, pp. 4696-4705.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.

[8] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and G. Michailidis, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in Adv. Neural Inf. Process. Syst. 32, 2019, pp. 13888-13899.

[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 6402-6413.

[10] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050-1059.

[11] Eric Wang and Heyu Wang, "QoE-Driven Reinforcement Learning for Joint Bitrate, Rebuffering, and TTFF Optimization in HLS/DASH", JACS, vol. 3, no. 2, pp. 50–59, Feb. 2023, doi: 10.69987/JACS.2023.30204.

[12] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, MIT Press, 2000, pp. 61-74.

[13] G. W. Brier, "Verification of forecasts expressed in terms of probability," Mon. Weather Rev., vol. 78, no. 1, pp. 1-3, 1950.

[14] A. Kumar, P. S. Liang, and T. Ma, "Verified uncertainty calibration," in Adv. Neural Inf. Process. Syst. 32, 2019, pp. 3792-3803.

[15] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in Adv. Neural Inf. Process. Syst. 31, 2018, pp. 7047-7058.

[16] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," arXiv:1701.06548, 2017.

[17] A. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 2151-2159.

[18] X. Geng, "Label distribution learning," IEEE Trans. Knowl. Data Eng., vol. 28, no. 7, pp. 1734-1748, 2016.

[19] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 41-50.

[20] Shenghan Lu and David Zhou, "LLM-Augmented Customer Representation Learning for Next-Purchase Prediction in Online Retail", JACS, vol. 3, no. 3, pp. 50–64, Mar. 2023, doi: 10.69987/JACS.2023.30305.