

References

- [1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.
- [2] Robert B Ash. *Information theory*. Courier Corporation, 2012.
- [3] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35:29205–29216, 2022.
- [4] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pages 2078–2091. PMLR, 2023.
- [5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.
- [6] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.
- [7] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020.
- [8] Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations*, 2022.
- [9] Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *The Twelfth International Conference on Learning Representations*.
- [10] Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [11] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008.
- [12] Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *International conference on machine learning*, pages 7596–7616. PMLR, 2023.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- [14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [16] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.
- [17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [18] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

- [19] Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors. *Advances in Neural Information Processing Systems*, 37:116987–117014, 2024.
- [20] Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.
- [21] Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.
- [22] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR, 2021.
- [23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [24] Gustaf Kylberg. *Kylberg texture dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and ..., 2011.
- [25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- [26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
- [27] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [28] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.
- [29] Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.
- [30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.
- [31] Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: theory and practice. In *International Conference on Machine Learning*, pages 26963–26989. PMLR, 2023.
- [32] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [33] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.
- [34] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.
- [35] Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [36] Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:84645–84668, 2024.
- [37] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

- 434 [38] Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das,
435 and Gregory Wornell. Are uncertainty quantification capabilities of evidential deep learning a
436 mirage? *Advances in Neural Information Processing Systems*, 37:107830–107864, 2024.
- 437 [39] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah
438 Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural
439 networks. *Advances in neural information processing systems*, 32, 2019.
- 440 [40] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on
441 evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*,
442 2021.
- 443 [41] Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep
444 learning. *arXiv preprint arXiv:2409.08754*, 2024.
- 445 [42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond
446 empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- 447 [43] Jingyang Zhang, Jingkan Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang,
448 Yiyu Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5:
449 Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.
- 450 [44] Qingyang Zhang, Qiuxuan Feng, Joey Tianyi Zhou, Yatao Bian, Qinghua Hu, and Changqing
451 Zhang. The best of both worlds: On the dilemma of out-of-distribution detection. *Advances in
452 Neural Information Processing Systems*, 37:69716–69746, 2024.
- 453 [45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A
454 10 million image database for scene recognition. *IEEE transactions on pattern analysis and
455 machine intelligence*, 40(6):1452–1464, 2017.

A Related work

A.1 Evidential Deep Learning

In classification tasks, Evidential Deep Learning (EDL) estimates predictive uncertainty by learning the parameters of a higher-order prior distribution over the categorical distribution, typically modeled as a Dirichlet distribution. This idea of modeling higher-order uncertainty has also been extended to regression tasks, where EDL approaches learn priors over the parameters of a Gaussian distribution—namely, its mean and variance—to capture uncertainty in continuous predictions [1, 8]. While our work focuses primarily on classification, readers interested in a broader perspective may refer to the comprehensive survey in [40]. In the classification setting, several prominent approaches have emerged. One foundational method is the EDL model proposed by Sensoy et al. [37], which incorporates the principles of Subjective Logic [20] to model class probabilities via a Dirichlet distribution. It introduces a KL divergence-based regularizer to discourage overconfident predictions on incorrect classes, thereby promoting better-calibrated uncertainty. Prior Networks (KL-PN) [28] explicitly learn a prior Dirichlet distribution to match a synthetic target distribution using forward KL divergence. These models rely on out-of-distribution (OOD) examples during training to enforce high uncertainty on unfamiliar inputs. This idea is further extended in Reverse KL Prior Networks (RKL-PN) [29], which instead maximize the reverse KL divergence between in-distribution and OOD predictions, encouraging sharper in-distribution certainty while maintaining uncertainty on OOD data. To move beyond the requirement of OOD data during training, Posterior Networks (PostNet) [7] leverage normalizing flows to learn a posterior over pseudo-counts, allowing flexible uncertainty modeling. Natural Posterior Networks (NatPN) [8] generalize this idea by placing the model within the exponential family framework, offering a unified approach to both classification and regression through Bayesian updates with learned latent densities. Recent advances focus on addressing overconfidence and improving robustness. \mathcal{I} -EDL [12] introduces a Fisher information-based regularizer to prevent the model from assigning excessive confidence, while R-EDL [10] modifies the prior and introduces constraints to avoid the degenerate behavior of collapsing variance in the Dirichlet distribution. H-EDL [35] further enriches the representational capacity by adopting hyper-opinions from subjective logic, thereby capturing second-order uncertainty. Lastly, DA-EDL [41] proposes to calibrate Dirichlet parameters using the local density of the feature space, enhancing reliability in sparse or uncertain regions. Together, these methods represent a growing body of work that refines evidential reasoning in neural networks, aiming for more robust and interpretable uncertainty estimation in classification tasks.

A.2 Critiques and Improvements of EDL Methods

Despite the notable progress of EDL models in uncertainty estimation, several recent works have raised critical concerns regarding their theoretical foundations and practical reliability. Bengs et al. [3] and Jurgens et al. [21] highlight a key issue: optimizing level-2 distributions via hard label supervision leads to a Dirac delta function without any regularization on the Dirichlet distribution. Bengs et al. [4] further argue that, under level-0 label supervision, virtually no proper scoring rules exist for level-2 distributions—even though neural networks have the theoretical capacity to approximate arbitrary distributions. While these studies effectively diagnose the limitations of EDL, they do not provide actionable insights or concrete solutions to overcome them. More recently, Shen et al. [38] offered an in-depth analysis of these critiques and proposed modeling epistemic uncertainty through model ensembles. While this method shows promising results, its major drawback lies in the computational cost of training multiple models—though inference can be performed using only a single distilled model. Overall, these analyses are all based on the assumption that supervision is provided in the form of level-0 hard labels. They overlook the implications of such fully certain supervision in the context of uncertainty-aware modeling. This assumption may itself be a root cause of overconfidence and misleading uncertainty estimates in EDL systems.

B List of Symbols

A list of symbols used in the main paper as well as in the following supplementary material, most of symbols keep same as [3] [4].

Table 4: Notation summary for the general, level-1, and level-2 learning settings.

General Symbols	
K	number of classes
\mathcal{X}	instance space
\mathcal{Y}	label space with hard labels $\{y_1, \dots, y_K\}$
\mathcal{D}	training data $\{(x^{(n)}, y^{(n)})\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$
P	data generating probability
$p(\cdot x)$	a conditional distribution on \mathcal{Y} , i.e., $p(y x)$, represents the probability of observing y given x
$\mathbb{P}(\mathcal{Y}), \mathbb{P}_1(\mathcal{Y})$	the set of probability distributions on \mathcal{Y}
Δ_K	the K -simplex, i.e., $\Delta_K := \{\theta = (\theta_1, \dots, \theta_K) \in [0, 1]^K \mid \ \theta\ _1 = 1\}$
$\theta = (\theta_1, \dots, \theta_K)^\top$	probability vector with K singletons
Level-1 Learning Setting	
\mathcal{H}_1	(level-1) hypothesis space consisting of hypothesis $h : \mathcal{X} \rightarrow \Delta_K$
L_1	loss function for a level-1 hypothesis, i.e., $L_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathcal{Y} \rightarrow \mathbb{R}$
$R(\cdot)$	risk or expected loss of a level-1 hypothesis (Eq. 3)
$\hat{R}_{\text{emp}}(\cdot)$	empirical loss of a level-1 hypothesis (Eq. 4)
\hat{h}	empirical risk minimiser, i.e., $\hat{h} = \arg \min_{h \in \mathcal{H}} \hat{R}_{\text{emp}}(h)$
h^*	true risk minimiser or Bayes predictor, i.e., $h^* = \arg \min_{h \in \mathcal{H}} R(h)$
Level-2 Learning Setting	
$\Delta_K^{(2)}$	the set of distributions on simplex Δ_K
$\mathbb{P}_2(\mathcal{Y})$	the set of distributions on $\mathbb{P}_1(\mathcal{Y})$ (the set of level-2 distributions)
\mathcal{H}_2	(level-2) hypothesis, i.e., a mapping $h : \mathcal{X} \rightarrow \Delta_K^{(2)}$
Q	probability distribution on $\mathbb{P}_1(\mathcal{Y})$, i.e., an element of $\mathbb{P}_2(\mathcal{Y})$
Q_{uni}	uniform distribution on $\mathbb{P}_1(\mathcal{Y})$ (an element of $\mathbb{P}_2(\mathcal{Y})$)
L_2	loss function for level-2 hypothesis, e.g., $L_2 : \mathbb{P}_2(\mathcal{Y}) \times (\cdot) \rightarrow \mathbb{R}_+$
$\hat{R}_{\text{emp}}^{(2)}(\cdot)$	empirical (level-2) loss of a level-2 hypothesis
$R^{(2)}(\cdot)$	(level-2) risk or expected loss of a level-2 hypothesis
Distributions	
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with location parameter μ and scale parameter $\sigma > 0$
$\text{Dir}(\alpha)$	Dirichlet distribution with parameter $\alpha \in \mathbb{R}_+^K$
δ_y	Dirac measure at $y \in \mathcal{Y}$ (i.e. δ_y is an element of $\mathbb{P}_1(\mathcal{Y})$)
δ_p	Dirac measure at $p \in \mathbb{P}_1(\mathcal{Y})$ (i.e., δ_p is an element of $\mathbb{P}_2(\mathcal{Y})$)
Entropy and Divergence	
$H(\cdot)$	Shannon Entropy of a categorical distribution
$\text{KL}(\cdot, \cdot)$	Kullback-Leibler divergence on $\mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y})$

507 C Proof of Theorem

508 **Theorem 1.** For any level-1 loss function $L_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathcal{Y} \rightarrow \mathbb{R}$ that satisfy $L_1(\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\alpha)} \mathbf{p}, \cdot) \leq$
509 $\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\alpha)} L_1(\mathbf{p}, \cdot)$, (i.e., as a convex function), such as Brier score and the log-loss in Eq. 2 empirical
510 risk minimizer of a level-2 prediction is always a Dirac measure $\delta_p \in \mathbb{P}_2(\mathcal{Y})$ and the expectation of
511 level-2 prediction is $\delta_y \in \mathbb{P}_1(\mathcal{Y})$.

512 *Proof.* Let the empirical risk of a level-2 prediction $Q \in \mathbb{P}_2(\mathcal{Y})$ as

$$\begin{aligned}
\hat{R}_{\text{emp}}^{(2)}(Q) &= \frac{1}{N} \sum_{n=1}^N L_2(Q, y^{(n)}) \\
&= \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathbf{p} \sim Q} L_1(\mathbf{p}, y^{(n)}) .
\end{aligned} \tag{18}$$

513 By assumption on the level-1 loss L_1 (i.e. convexity), it holds that

$$\hat{R}_{\text{emp}}^{(2)}(Q) \geq \frac{1}{N} \sum_{n=1}^N L_1 \left(\mathbb{E}_{\mathbf{p} \sim Q}[\mathbf{p}], y^{(n)} \right). \quad (19)$$

514 Let $\tilde{Q}^{(N)}$ be the minimiser over all $Q \in \Delta_K^{(2)}$ of the right-hand side, then $\tilde{\mathbf{p}}^{(N)} = \mathbb{E}_{\mathbf{p} \sim \tilde{Q}^{(N)}}[\mathbf{p}]$ is an
 515 element in Δ_K . Define $\hat{Q}^{(N)} = \delta_{\tilde{\mathbf{p}}^{(N)}}$ and note that $\mathbb{E}_{\mathbf{p} \sim \hat{Q}^{(N)}}[\mathbf{p}] = \tilde{\mathbf{p}}^{(N)}$. Then,

$$\begin{aligned} \hat{R}_{\text{emp}}^{(2)}(\hat{Q}^{(N)}) &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathbf{p} \sim \hat{Q}^{(N)}} L_1(\mathbf{p}, y^{(n)}) \\ &= \frac{1}{N} \sum_{n=1}^N L_1(\tilde{\mathbf{p}}^{(N)}, y^{(n)}) \\ &= \frac{1}{N} \sum_{n=1}^N L_1 \left(\mathbb{E}_{\mathbf{p} \sim \tilde{Q}^{(N)}}[\mathbf{p}], y^{(n)} \right). \end{aligned} \quad (20)$$

516 This proves that the empirical level-2 risk is minimized by a Dirac distribution over a single level-1
 517 prediction, i.e., $\hat{Q}^{(N)} = \delta_{\tilde{\mathbf{p}}^{(N)}}$, implying vanishing epistemic uncertainty. We now show that the
 518 corresponding level-1 prediction also collapses to a Dirac measure, indicating vanishing aleatoric
 519 uncertainty. Consider the empirical level-1 risk:

$$\hat{R}_{\text{emp}}^{(1)}(\mathbf{p}) = \frac{1}{N} \sum_{n=1}^N L_1(\mathbf{p}, y^{(n)}). \quad (21)$$

520 For any strictly proper loss function L_1 (e.g., Brier score, log-loss), it is uniquely minimized when
 521 $\mathbf{p} = \delta_{y^{(n)}}$, i.e., the one-hot encoding of the ground-truth label. That is,

$$\arg \min_{\mathbf{p} \in \Delta_K} L_1(\mathbf{p}, y^{(n)}) = \delta_{y^{(n)}}, \quad \text{with} \quad L_1(\delta_{y^{(n)}}, y^{(n)}) = 0. \quad (22)$$

522 Hence, the optimal level-1 predictor $\tilde{\boldsymbol{\theta}}^{(N)}$ that minimizes the empirical risk is

$$\tilde{\mathbf{p}}^{(N)} = \delta_{y^{(n)}}, \quad \text{for all } n. \quad (23)$$

523 It follows that the expected level-1 prediction under the optimal level-2 distribution is

$$\mathbb{E}_{\mathbf{p} \sim \hat{Q}^{(N)}} \mathbf{p} = \delta_{y^{(n)}}, \quad (24)$$

524 i.e., a one-hot distribution that assigns all probability mass to the ground-truth class. This indicates
 525 that aleatoric uncertainty also vanishes. \square

526 Therefore, the empirical level-2 risk is minimized by a Dirac measure over a level-1 Dirac prediction

$$\hat{Q}^{(N)} = \delta_{\delta_{y^{(n)}}}. \quad (25)$$

527 This implies that:

- 528 • **Epistemic uncertainty vanishes**, since Q is a Dirac measure.
- 529 • **Aleatoric uncertainty vanishes**, since the expected level-1 prediction under Q is a one-hot
 530 vector.

531 This highlights a critical degeneracy of empirical risk minimization with strictly proper convex losses
 532 in the level-2 setting: it collapses all predictive uncertainty, providing no representation of uncertainty
 533 despite operating in a distribution-over-distributions framework.

534 **Proposition 1.** *Under the assumptions of Theorem 1 empirical risk minimization of level-2 prediction*
 535 *inevitably yields degenerate distributions $\delta_{\mathbf{p}} \in \mathbb{P}_2(\mathcal{Y})$ and the expectation of level-2 prediction is*
 536 *$\delta_y \in \mathbb{P}_1(\mathcal{Y})$. As a result, the model fails to provide any meaningful or disentangled representation of*
 537 *aleatoric or epistemic uncertainty.*

538 *Proof.* Assume that the optimal strategy under ERM is to collapse the Dirichlet distribution to a delta
 539 distribution centered on the one-hot vector δ_y , i.e., $\text{Dir}(\alpha) \rightarrow \delta_{\delta_y}$ as in Theorem 1. This degeneracy
 540 has consequences for uncertainty estimation. Consider the standard decomposition of predictive
 541 uncertainty in Dirichlet-based models as D.1 we have

$$\text{Total Uncertainty (EU)} = H [\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\alpha)} [p(y|\mathbf{p})]], \quad (26)$$

$$\text{Aleatoric Uncertainty (AU)} = \mathbb{E}_{\mathbf{p}} [H[p(y|\mathbf{p})]], \quad (27)$$

$$\text{Epistemic Uncertainty (EU)} = \text{TU} - \text{AU}. \quad (28)$$

544 When the Dirichlet degenerates to δ_{δ_y} , both the expected predictive distribution and the samples from
 545 $\text{Dir}(\alpha)$ are deterministic, yielding

$$\text{TU} \rightarrow 0, \quad \text{AU} \rightarrow 0, \quad \text{EU} \rightarrow 0. \quad (29)$$

546 Thus, the model expresses neither AU nor EU, regardless of the true nature of the data distribution.
 547 Consequently, the level-2 model fails to provide any meaningful or disentangled representation of
 548 aleatoric or epistemic uncertainty. \square

549 **Theorem 2.** Let the ground-truth level-1 label be denoted as $\mathbf{p}^*(\mathbf{x})$, and let the observed level-0
 550 one-hot label $\delta_y(\mathbf{x})$ be a noisy realization of $\mathbf{p}^*(\mathbf{x})$ perturbed by input-dependent label noise $\mu(\mathbf{x})$

$$\delta_y(\mathbf{x}) = \mathbf{p}^*(\mathbf{x}) + \mu(\mathbf{x}) \quad \text{where} \quad \mu(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (30)$$

551 Then, the test risk admits the following lower bound under mild regularity conditions

$$R(\hat{h}; P) \geq C\sigma^2, \quad (31)$$

552 where C depends on the trace of the Hessian matrix of the loss function w.r.t. \mathbf{p} . Then, for the level-1
 553 label with strong mixing, the bound can be tightened as

$$R(\hat{h}; P) \geq C'\sigma^2, \quad (32)$$

554 where $C'/C \approx \frac{1}{2\beta+1} + \frac{1}{2} < 1$ ($\forall \beta \gg 1/2$), indicating a reduced sensitivity of the test risk to
 555 input-dependent noise.

556 *Proof.* We suppose the label noise μ follow an isotropic Gaussian distribution as \mathcal{I} -EDL [12]

$$\mu \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (33)$$

557 Then, even if the optimization loss the $R(\hat{h}; \mathcal{D})$ is minimized (or approaches zero), the population
 558 loss $R(\hat{h}; P)$ will have an irreducible component that is at least on the order of σ^2 . As we assume
 559 that the training labels y are generated from the true labels \mathbf{p}^* with added noise

$$\delta_y(\mathbf{x}) = \mathbf{p}^*(\mathbf{x}) + \mu(\mathbf{x}) \quad (34)$$

560 where $\mu(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. The expected test loss can be expressed as

$$R(\hat{h}; P) := \mathbb{E}_{(\mathbf{x}, y) \sim P} [L_2(\hat{h}(\mathbf{x}), y)]. \quad (35)$$

561 Since the label itself is affected by noise, we can decompose the expectation as:

$$\mathbb{E}[L_2(\hat{h}(\mathbf{x}), \delta_y)] = \mathbb{E}[L_2(\hat{h}(\mathbf{x}), \mathbf{p}^* + \mu)]. \quad (36)$$

562 Using a second-order Taylor expansion to approximate the loss function

$$\ell(\hat{h}(\mathbf{x}), \mathbf{p}^* + \mu) \approx L_2(\hat{h}(\mathbf{x}), \mathbf{p}^*) + \langle \nabla L_2, \mu \rangle + \frac{1}{2} \mu^T \mathbf{H} \mu \quad (37)$$

563 where \mathbf{H} represents the Hessian matrix of the loss function $L_2(\hat{h}(\mathbf{x}), \mathbf{p}^*)$ w.r.t. \mathbf{p}^* , defined as

$$\mathbf{H} = \nabla^2 L_2(\hat{h}(\mathbf{x}), \mathbf{p}^*) \quad (38)$$

564 and $\langle \nabla L_2, \mu \rangle$ is the inner product between the gradient of the loss function and the noise vector μ

$$\langle \nabla L_2, \mu \rangle = \sum_k^K \frac{\partial L_2}{\partial \mu_k} \mu_k. \quad (39)$$

Since the noise μ follows a zero-mean Gaussian distribution, the expectation of the first-order term vanishes.

$$\mathbb{E}[\langle \nabla L_2, \mu \rangle] = 0 \quad (40)$$

while the expectation of the second-order term is given by the noise covariance

$$\mathbb{E}[\mu^T H \mu] = \sigma^2 \text{Tr}(H). \quad (41)$$

Thus, the lower bound of the test loss can be approximated as

$$R(\hat{h}; \mathcal{P}) \geq C\sigma^2 \quad (42)$$

where C depends on the trace of the Hessian matrix. We then show that incorporating VRM leads to a lower test risk. Let the original label noise $\mu^{(n)}, \mu^{(m)} \sim \mathcal{N}(0, \sigma^2 I)$ be i.i.d. After vicinal interpolation, the noise in vicinal labels becomes

$$\tilde{\mu} = \lambda \mu^{(n)} + (1 - \lambda) \mu^{(m)} \quad (43)$$

with variance

$$\mathbb{E}\|\tilde{\mu}\|^2 = \lambda^2 \sigma^2 + (1 - \lambda)^2 \sigma^2 = \sigma^2 [\lambda^2 + (1 - \lambda)^2]. \quad (44)$$

When $\lambda \sim \text{Beta}(\beta, \beta)$, the expected variance is

$$\mathbb{E}_\lambda [\lambda^2 + (1 - \lambda)^2] = 2\mathbb{E}[\lambda^2] - 2\mathbb{E}[\lambda] + 1. \quad (45)$$

Using properties of Beta distribution $\mathbb{E}[\lambda] = \frac{1}{2}$ and $\text{Var}(\lambda) = \frac{1}{4(2\beta+1)}$, we obtain

$$\mathbb{E}[\lambda^2] = \text{Var}(\lambda) + (\mathbb{E}[\lambda])^2 = \frac{1}{4(2\beta+1)} + \frac{1}{4}. \quad (46)$$

Substituting yields

$$\mathbb{E}_\lambda [\lambda^2 + (1 - \lambda)^2] = \frac{1}{2\beta+1} + \frac{1}{2} < 1 \quad (\forall \beta \gg 1/2). \quad (47)$$

Thus, the effective noise variance after Mixup is $k\sigma^2$, where $k = \frac{1}{2\beta+1} + \frac{1}{2} < 1$, significantly lower than the original σ^2 . Substituting into the theorem's lower bound gives

$$R(\hat{h}; P) \geq C \cdot k\sigma^2 < C\sigma^2 \quad (48)$$

Although distribution of the noise μ is unknown; and assumptions about it is a modeling question, most statistical methods rely on certain mathematical conditions, known as regularity assumptions, to ensure their validity. In our proof, i.e., we assume that μ follows a additive gaussian noise. \square

Theorem 3. Let λ be the mixing hyperparameter defined in Eq. 12. Consider the optimization of the Dirichlet parameters α in Eq. 13. For samples where $\alpha_k \leq \frac{1}{K} \sum_{j=1}^K \alpha_j$ with low belief assigned to the ground-truth k class, the following properties hold

- The update to the Dirichlet concentration for the ground-truth class $\Delta\alpha_k$ increases monotonically with λ .
- The updates to the Dirichlet concentrations for the non-ground-truth classes $\Delta\alpha_{j \neq k}$ decrease monotonically with λ .
- The total increase in Dirichlet concentration, denoted ΔS , increases monotonically with λ .

Proof. Let k denotes the index of ground-truth class. By optimizing the objective function in Eq. 13 with gradient descent, the update of α_j with single gradient update can be denoted as

$$\Delta\alpha_j := -\eta \frac{\partial \mathcal{L}_{\text{edl}}}{\alpha_j}, \quad (49)$$

where η is the learning rate. Let j denote the index of class, we have

$$\frac{\partial \mathcal{L}_{\text{edl}}}{\alpha_j} = \tilde{y}_j [\psi_1(S) - \psi_1(\alpha_j)] \quad (50)$$

where ψ_1 is the trigamma function, which is a positive, monotonic decreasing and strictly convex function. Then, we have the accumulative updates of α_j with T -steps as

$$\Delta\alpha_j = \eta \sum_{t=1}^T \tilde{y}_j \left[\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right]. \quad (51)$$

As the vicinal label is obtained by $\tilde{\mathbf{y}} = \lambda \mathbf{y}^{(n)} + (1 - \lambda) \cdot [\frac{1}{K}, \dots, \frac{1}{K}]$, we can also express the smoothed target labels explicitly as

$$\tilde{y}_k = \lambda + \frac{1 - \lambda}{K}, \quad \tilde{y}_j = \frac{1 - \lambda}{K}, \quad (52)$$

where k denotes the index of ground-truth class. By substituting Eq. 52 into Eq. 51, we have

$$\Delta\alpha_k = \eta \sum_{t=1}^T \left(\lambda + \frac{1 - \lambda}{K} \right) \left[\psi_1(\alpha_k^{(t)}) - \psi_1(S^{(t)}) \right]. \quad (53)$$

and

$$\Delta\alpha_j = \eta \sum_{t=1}^T \frac{1 - \lambda}{K} \left[\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right]. \quad (54)$$

and

$$\begin{aligned} \Delta S &= \eta \sum_{t=1}^T \left[\left(\lambda + \frac{1 - \lambda}{K} \right) \left(\psi_1(\alpha_k^{(t)}) - \psi_1(S^{(t)}) \right) + \sum_{j \neq k} \frac{1 - \lambda}{K} \left(\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right) \right] \\ &= \eta \sum_{t=1}^T \left[\lambda \left(\psi_1(\alpha_k^{(t)}) - \psi_1(S^{(t)}) \right) + \frac{1 - \lambda}{K} \sum_{j=1}^K \left(\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right) \right] \end{aligned} \quad (55)$$

Then, the total update of Dirichlet concentration ΔS can also be expressed as

$$\begin{aligned} \Delta S &= \sum_{j=1}^K \Delta\alpha_j = \eta \sum_{t=1}^T \sum_{j=1}^K \tilde{y}_j \left[\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right] \\ &= \eta \sum_{t=1}^T \left[\underbrace{\sum_{j=1}^K \tilde{y}_j \psi_1(\alpha_j^{(t)})}_{\text{a convex combination of the } \psi_1(\alpha_j)} - \psi_1(S^{(t)}) \right] \\ &\geq \eta \sum_{t=1}^T \left[\psi_1 \left(\sum_{j=1}^K \tilde{y}_j \alpha_j^{(t)} \right) - \psi_1(S^{(t)}) \right] \geq 0 \end{aligned} \quad (56)$$

To analyze how λ affects $\Delta\alpha_k$, $\Delta\alpha_j$, and ΔS , consider the derivatives

$$\frac{\partial \Delta\alpha_k}{\partial \lambda} = \eta \sum_{t=1}^T \left(1 - \frac{1}{K} \right) \left[\psi_1(\alpha_k^{(t)}) - \psi_1(S^{(t)}) \right] > 0, \quad (57)$$

$$\frac{\partial \Delta\alpha_j}{\partial \lambda} = -\eta \sum_{t=1}^T \frac{1}{K} \left[\psi_1(\alpha_j^{(t)}) - \psi_1(S^{(t)}) \right] < 0, \quad (58)$$

and

$$\begin{aligned} \frac{\partial \Delta S}{\partial \lambda} &= \eta \sum_{t=1}^T \left[\psi_1(\alpha_k^{(t)}) - \frac{1}{K} \sum_{j=1}^K \psi_1(\alpha_j^{(t)}) \right] \\ &\geq \eta \sum_{t=1}^T \left[\psi_1(\alpha_k^{(t)}) - \psi_1 \left(\frac{1}{K} \sum_{j=1}^K \alpha_j^{(t)} \right) \right] \end{aligned} \quad (59)$$

For samples where $\alpha_k \leq \frac{1}{K} \sum_{j=1}^K \alpha_j$ —which typically correspond to uncertain predictions with low belief assigned to the ground-truth class in the early stages of training—we have

$$\frac{\partial \Delta S}{\partial \lambda} \geq 0. \quad (60)$$

Therefore, decreasing λ (i.e. with stronger noise mixed) for such samples can effectively suppress the growth of their Dirichlet concentration parameters. This results in a more dispersed (i.e., less confident) predictive distribution, thereby promoting higher epistemic uncertainty. \square

D Uncertainty Measures

D.1 Uncertainty Decomposition in Dirichlet-Based Models

A fundamental identity in information theory is that the Shannon entropy of a random variable X can be additively decomposed into the mutual information between X and Y , and the conditional entropy of X given Y [2]:

$$H(X) = I(X; Y) + H(X | Y) \quad (61)$$

Follow this idea, Prior Networks [28] propose a method to explicitly model and decompose predictive total uncertainty into two components: *data uncertainty* (aleatoric uncertainty) and *distributional uncertainty* (epistemic uncertainty). The total uncertainty in the prediction is measured by the Shannon entropy of the expected categorical distribution conditioned

$$H[p(y|\mathbf{p})] = \mathbb{E}_{\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})}[p(y | \mathbf{p})] = - \sum_{j=1}^K \frac{\alpha_j}{S} \log \frac{\alpha_j}{S}, \quad (62)$$

Aleatoric uncertainty (or data uncertainty) corresponds to the expected entropy of the categorical distributions sampled from the Dirichlet prior, commonly referred to as the *conditional entropy*

$$\begin{aligned} \mathbb{E}_{\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})}[H[p(y | \mathbf{p})]] &= \mathbb{E}_{\mathbf{p}} \left[- \sum_{j=1}^K p_j \log p_j \right] \\ &= - \sum_{j=1}^K \frac{\alpha_j}{S} (\psi(\alpha_j + 1) - \psi(S + 1)) \\ &= \psi(S + 1) - \sum_{j=1}^K \frac{\alpha_j}{S} \psi(\alpha_j + 1) \end{aligned} \quad (63)$$

where $\psi(\cdot)$ is the digamma function. For epistemic uncertainty, it is measured by the mutual information between predictions and the Dirichlet parameters as

$$\text{MI}(y, \mathbf{p}) = H_{\text{total}} - \mathbb{E}_{\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})}[H[p(y | \mathbf{p})]]. \quad (64)$$

This mutual information quantifies how much of the total uncertainty arises from uncertainty in the model parameters (i.e., distribution over categorical distributions), and thus reflects *epistemic uncertainty*.

$$\begin{aligned} \underbrace{\text{MI}[y, \mathbf{p}]}_{\text{Epistemic Uncertainty}} &\approx \underbrace{H[\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})}[p(y|\mathbf{p})]]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})}[H[p(y|\mathbf{p})]]}_{\text{Aleatoric Uncertainty}} \\ &= - \sum_{j=1}^K \frac{\alpha_j}{S} \ln \frac{\alpha_j}{S} + \sum_{j=1}^K \frac{\alpha_j}{S} (\psi(\alpha_j + 1) - \psi(S + 1)) \\ &= - \sum_{j=1}^K \frac{\alpha_j}{S} \left(\ln \frac{\alpha_j}{S} - \psi(\alpha_j + 1) + \psi(S + 1) \right). \end{aligned} \quad (65)$$

624 D.2 Differential Entropy

625 The *differential entropy* is defined as

$$\text{ENT}(\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha})) = - \int_{\Delta_K} \text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) \log \text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p}, \quad (66)$$

626 where Δ_K denotes the probability simplex. The closed-form expression is given by

$$\text{ENT}(\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha})) = \log B(\boldsymbol{\alpha}) + (S - K)\psi(S) - \sum_{j=1}^K (\alpha_j - 1)\psi(\alpha_j), \quad (67)$$

627 Differential entropy is also a prevalent measure of epistemic uncertainty, where a lower value indicates
628 that the model yields a sharper distribution, and a higher value means a more uniform Dirichlet
629 distribution.

630 D.3 Vacuity of Evidence

631 EDL [37], RED [31], \mathcal{I} -EDL [12], R-EDL [10], and H-EDL [36], which are grounded in Subjective
632 Logic [20] and Dempster-Shafer Theory [11], represent uncertainty using evidence mass values.
633 Subjective Logic [20] provides a principled framework for modeling predictive uncertainty by
634 interpreting the output of a neural network as an *opinion*—a structured representation of uncertainty
635 over a discrete set of classes. Unlike conventional classifiers that output categorical probabilities,
636 EDL models produce non-negative evidence values $\mathbf{e} = [e_1, e_2, \dots, e_K]$ for each of the K classes.
637 These evidence values parameterize a Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$, where $\alpha_j = e_j + 1$. In Subjective
638 Logic, an opinion over a finite domain is characterized by three components: the belief mass b_j , the
639 base rate a_j , and the uncertainty mass u , satisfying:

$$b_j + u \cdot a_j = \mathbb{E}[p_j], \quad \text{and} \quad \sum_{j=1}^K b_j + u = 1 \quad (68)$$

640 where p_j denotes the probability assigned to class j . These quantities relate to the Dirichlet parameters
641 as follows: The belief mass b_k is proportional to the evidence for class k :

$$b_k = \frac{e_k}{S}, \quad \text{where } S = \sum_{j=1}^K (e_j + 1) = \sum_{j=1}^K \alpha_j \quad (69)$$

642 The base rate a_k is typically assumed to be uniform, i.e., $a_k = 1/K$. The uncertainty mass u is
643 defined as:

$$u = \frac{K}{\sum_{j=1}^K \alpha_j} = \frac{K}{S} \quad (70)$$

644 This uncertainty mass u is referred to as *vacuity* in EDL literature, and it quantifies the degree
645 of epistemic uncertainty due to a lack of evidence. When the total evidence is low (e.g., under
646 out-of-distribution or ambiguous inputs), S becomes small and vacuity u approaches 1, indicating
647 that the model abstains from committing belief to any specific class. Conversely, high total evidence
648 yields a low vacuity, reflecting confident predictions based on strong feature-based support. This
649 opinion-based interpretation highlights the epistemic nature of uncertainty in EDL and differentiates
650 it from aleatoric uncertainty captured by distributional spread in conventional probabilistic models.

651 E Experimental Details

652 E.1 Implementation Details

653 Since different baseline methods involve distinct activation functions and regularization terms, we
654 provide detailed implementation settings below.

655 **EDLs based on Subjective Logic.** For EDL [37], we adopt the *barrier score*, i.e., the mean squared
656 error (MSE) loss. For \mathcal{I} -EDL [12], we follow their original paper and use the Fisher-MSE loss, where
657 the Fisher information regularization weight is set to 0.05. The activation function is `Softplus`, as

specified in their implementation. For R-EDL [10], we follow the settings in the original paper and set the prior strength to 0.8 for the CIFAR datasets. The loss function used is the MSE loss variant without the variance minimization term. For all three methods (EDL, \mathcal{I} -EDL, and R-EDL), the KL divergence term which aims to remove misleading evidence with an annealing weight schedule of $\lambda_t = \min(\text{epoch_idx}/10, 1)$. The KL divergence term which is used to regularize the predicted Dirichlet distribution by encouraging it to stay close to a non-informative prior for incorrect classes, typically $\text{Dir}(\mathbf{p} \mid \mathbf{1})$, where each class has a concentration parameter of 1. The KL divergence between the predicted Dirichlet distribution $\text{Dir}(\mathbf{p} \mid \bar{\alpha})$ and the uniform Dirichlet prior

$$\begin{aligned}\mathcal{L}_{\text{KL}} &= \text{KL}[\text{Dir}(\mathbf{p} \mid \bar{\alpha}) \parallel \text{Dir}(\mathbf{p} \mid \mathbf{1})] \\ &= \log \left(\frac{\Gamma \left(\sum_{j=1}^K \alpha_j \right)}{\prod_{j=1}^K \Gamma(\alpha_j)} \right) + \sum_{j=1}^K (\alpha_j - 1) \left[\psi(\alpha_j) - \psi \left(\sum_{j=1}^K \alpha_j \right) \right]\end{aligned}\quad (71)$$

where $\bar{\alpha} := \mathbf{y} + (1 - \mathbf{y}) \odot \alpha$ represents a modified Dirichlet parameter vector where the target class value is set to 1, $\Gamma(\cdot)$ is the gamma function, $\psi(\cdot)$ is the digamma function. Then, the total objective function is

$$\mathcal{L} = \mathcal{L}_{\text{EDL}} + \lambda_t \cdot \mathcal{L}_{\text{KL}} \quad (72)$$

PriorNets. For KL-PN [28] and RKL-PN [29], we set the target class Dirichlet concentration parameter to $\alpha_k = 200$. Since both methods require out-of-distribution (OOD) samples during training to constrain the predicted Dirichlet distributions, we follow the setup in [12, 9] and use random noise as the OOD dataset to ensure a fair comparison.

Our method. For our method, the non-negative activation function σ is set to `SoftPlus` for CIFAR-10. For CIFAR-100, due to the large zero-evidence regions observed in prior work [31], we warm up the model using an `Exponential` activation for the first 10 epochs to move away from the regions, and then switch to `SoftPlus` activation.

F Discussions

F.1 Why do some methods perform poorly on CIFAR-100?

For models like EDLs [37, 10] and PriorNets [28, 29] that require Dirichlet concentrations for incorrect classes to approach zero, we observe that they struggle to converge when the number of classes is large (e.g., $K = 100$). Since the original papers do not provide CIFAR-100 experimental settings, we adopt the same configurations as used for CIFAR-10, which may limit their performance.

F.2 Social Impact

Our work addresses the challenges of uncertainty estimation, out-of-distribution (OOD) detection and out-of-distribution (OOD) generalization, which are critical for ensuring the safety, reliability, and fairness of machine learning systems in real-world applications. By improving the model’s ability to recognize and appropriately respond to unfamiliar or ambiguous inputs, our methods contribute to reducing the risk of overconfident mispredictions in high-stakes domains such as healthcare, autonomous driving, and finance. These advances have the potential to increase trust in AI systems and support more responsible deployment practices. Moreover, enhanced OOD generalization may help mitigate performance disparities when models are applied across diverse populations and settings.

G Uncertainty distribution

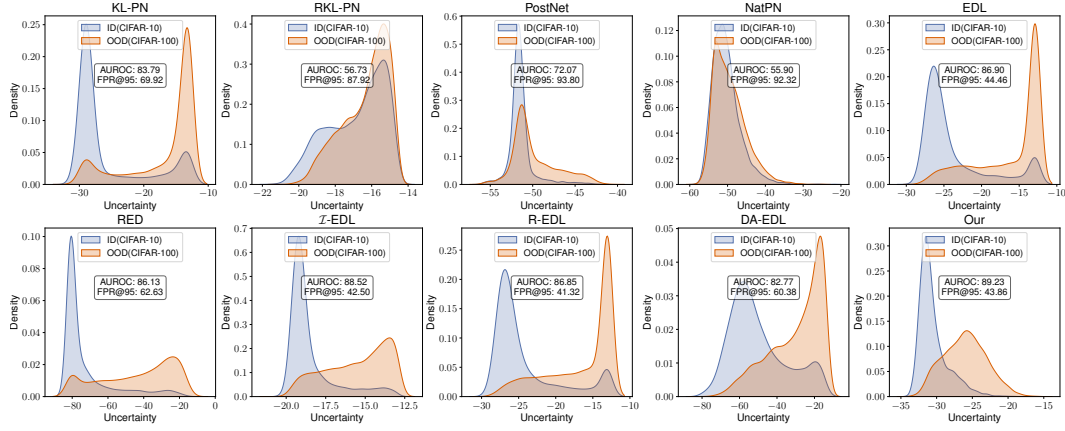


Figure 4: Differential entropy distribution on CIFAR10 (ID) vs CIFAR-100 (OOD).

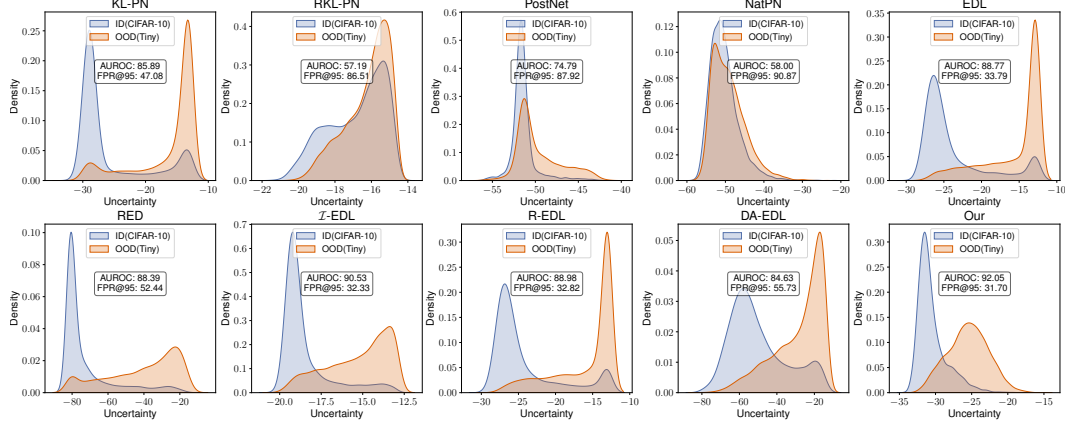


Figure 5: Differential entropy distribution on CIFAR10 (ID) vs Tiny-ImageNet (OOD).

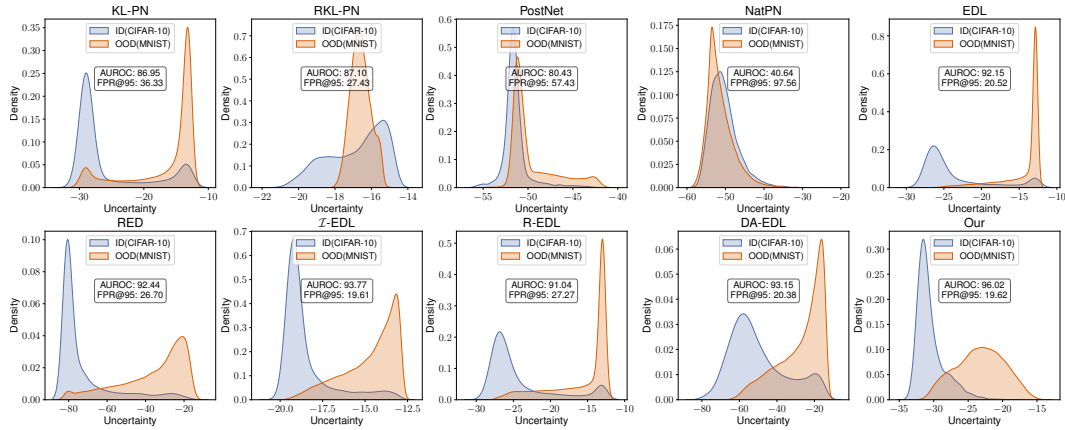


Figure 6: Differential entropy distribution on CIFAR10 (ID) vs MNIST (OOD).

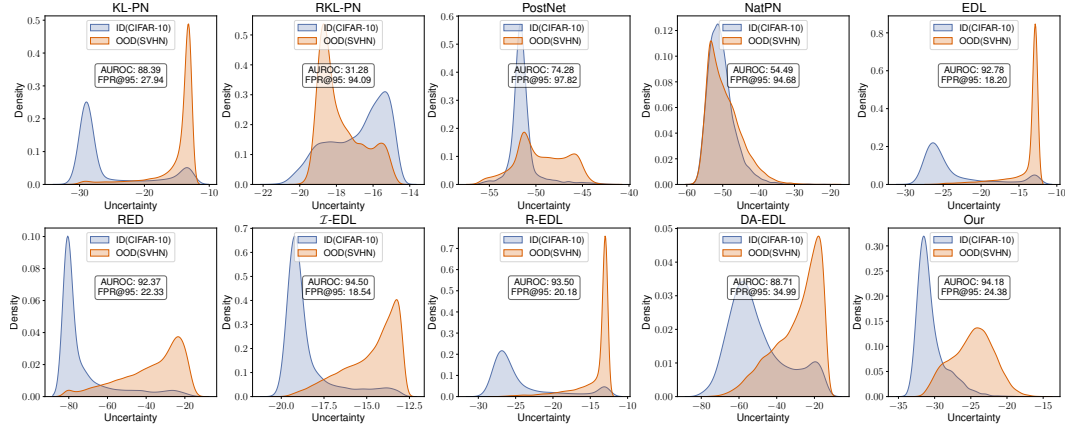


Figure 7: Differential entropy distribution on CIFAR10 (ID) vs SVHN (OOD).

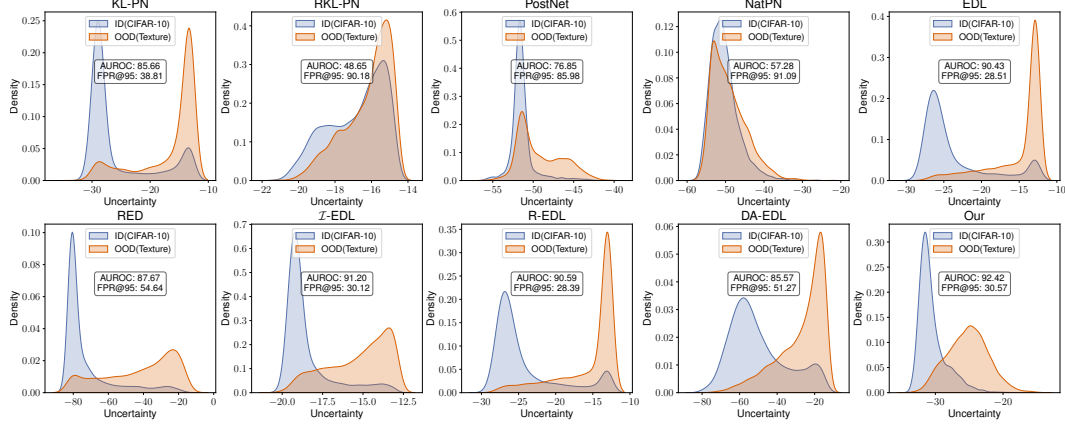


Figure 8: Differential entropy distribution on CIFAR10 (ID) vs Texture (OOD).

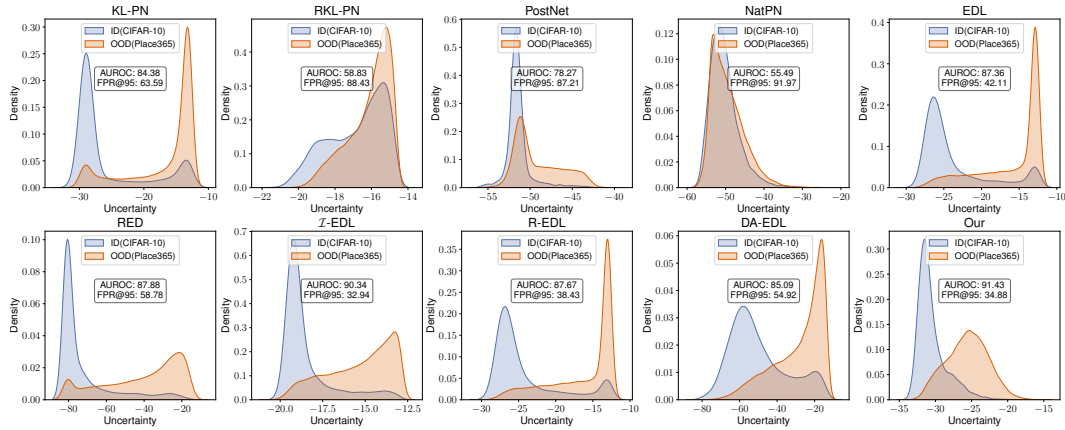


Figure 9: Differential entropy distribution on CIFAR10 (ID) vs Place365 (OOD).

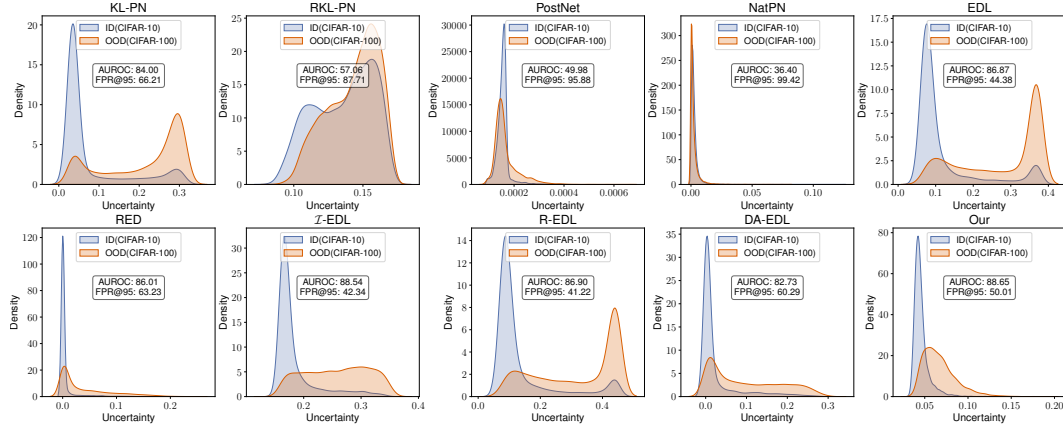


Figure 10: Mutual information distribution on CIFAR10 (ID) vs CIFAR-100 (OOD).

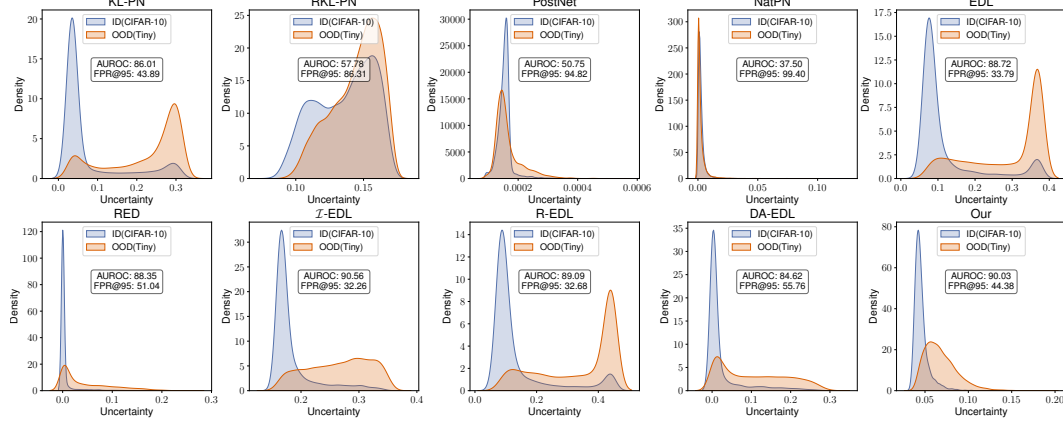


Figure 11: Mutual information distribution on CIFAR10 (ID) vs Tiny-ImageNet (OOD).

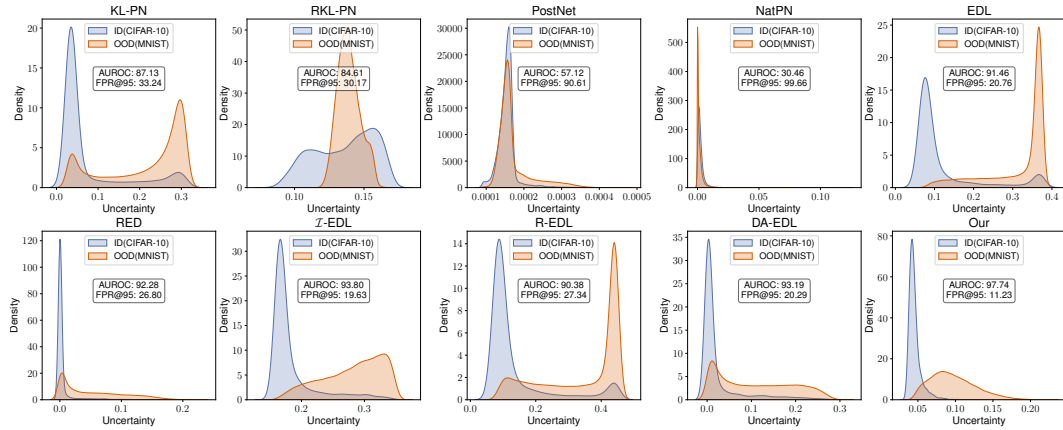


Figure 12: Mutual information distribution on CIFAR10 (ID) vs MNIST (OOD).

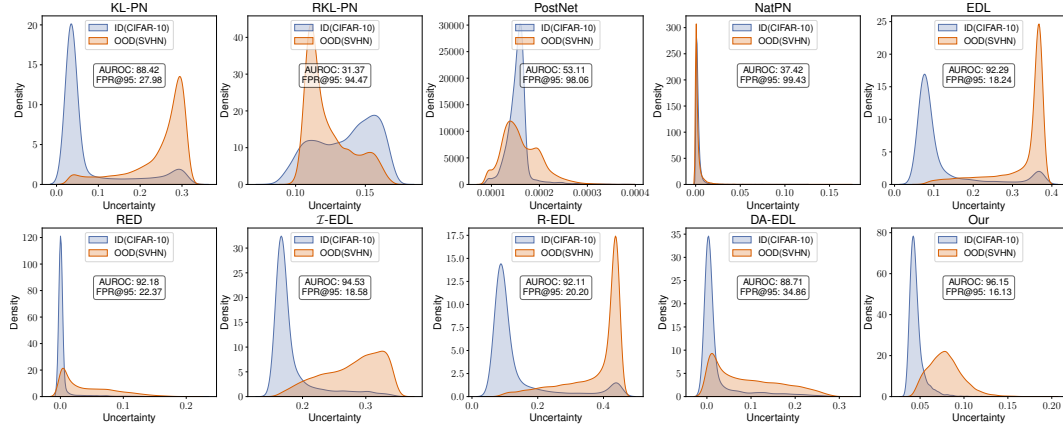


Figure 13: Mutual information distribution on CIFAR10 (ID) vs SVHN (OOD).

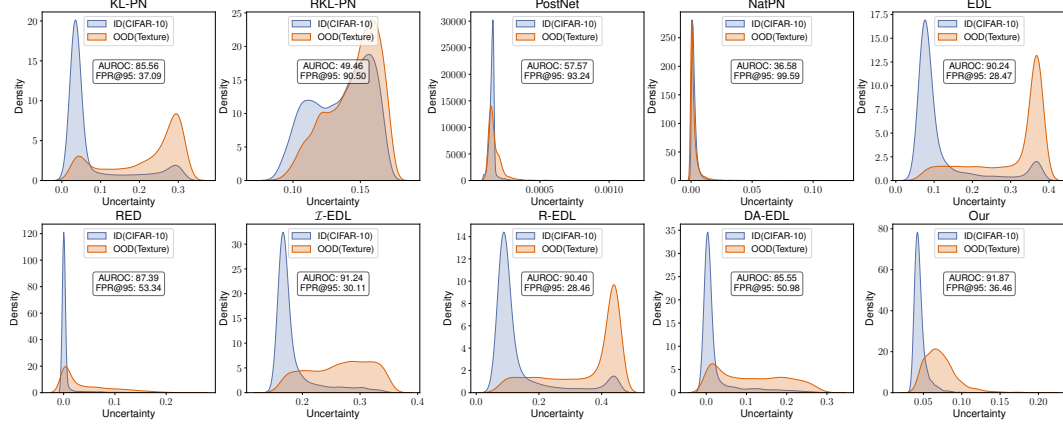


Figure 14: Mutual information distribution on CIFAR10 (ID) vs Texture (OOD).

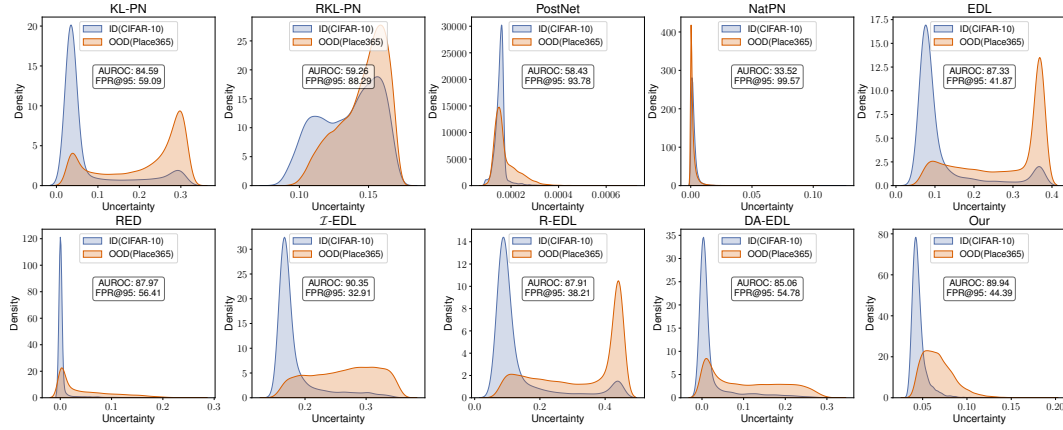


Figure 15: Mutual information distribution on CIFAR10 (ID) vs Place365 (OOD).

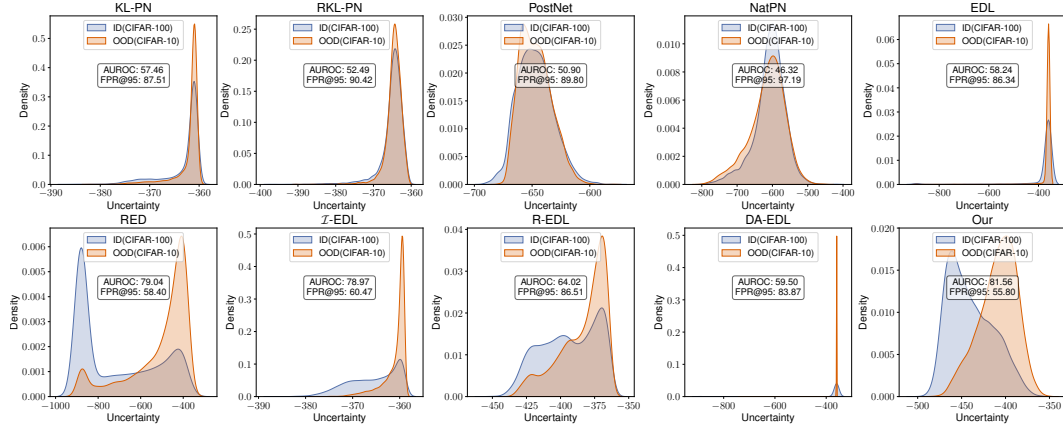


Figure 16: Differential entropy distribution on CIFAR100 (ID) vs CIFAR-10 (OOD).

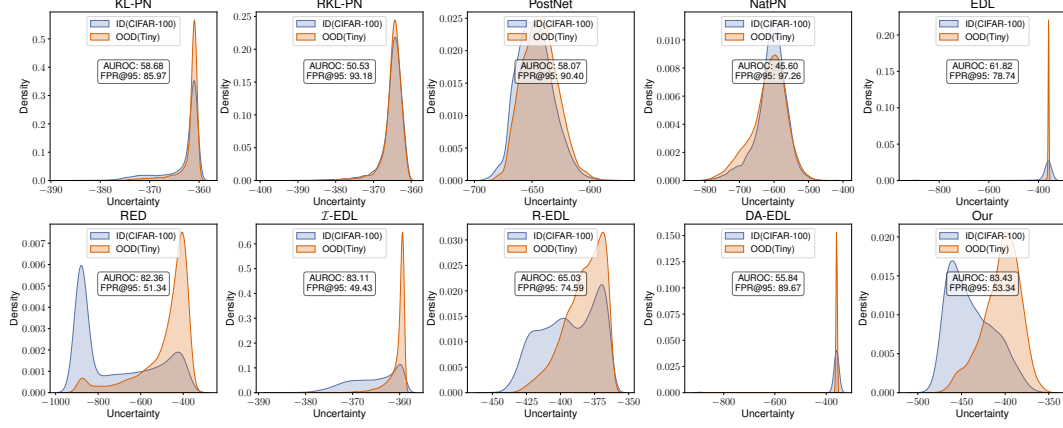


Figure 17: Differential entropy distribution on CIFAR100 (ID) vs Tiny-ImageNet (OOD).

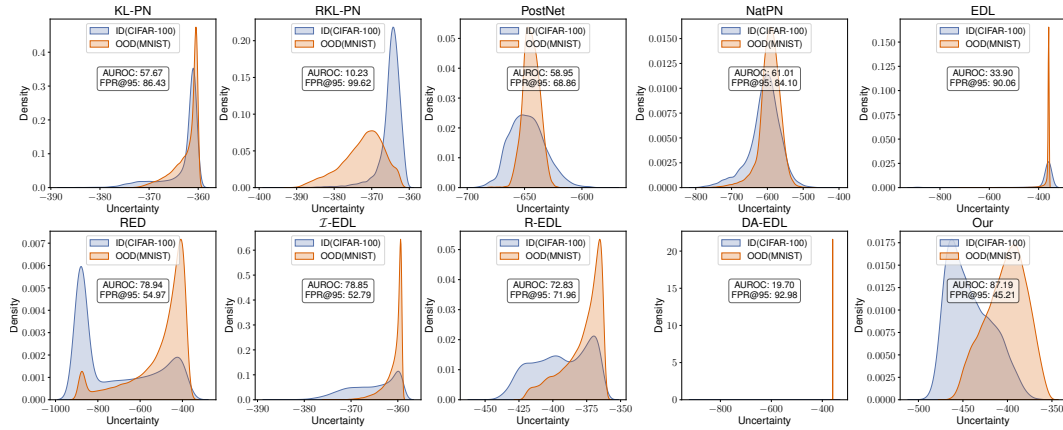


Figure 18: Differential entropy distribution on CIFAR100 (ID) vs MNIST (OOD).

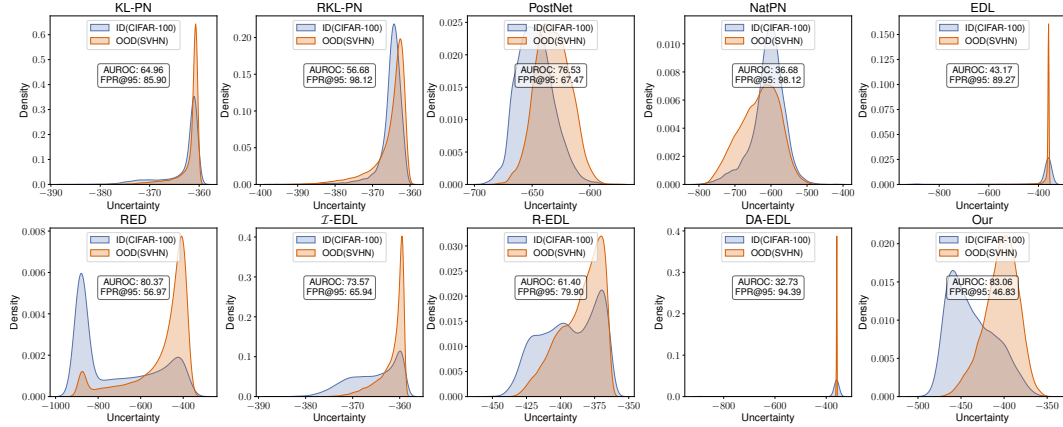


Figure 19: Differential entropy distribution on CIFAR100 (ID) vs SVHN (OOD).

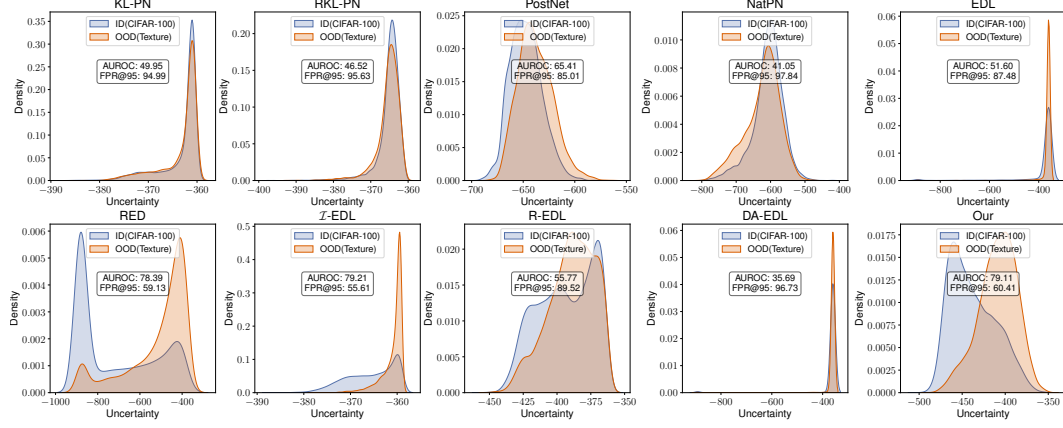


Figure 20: Differential entropy distribution on CIFAR100 (ID) vs Texture (OOD).

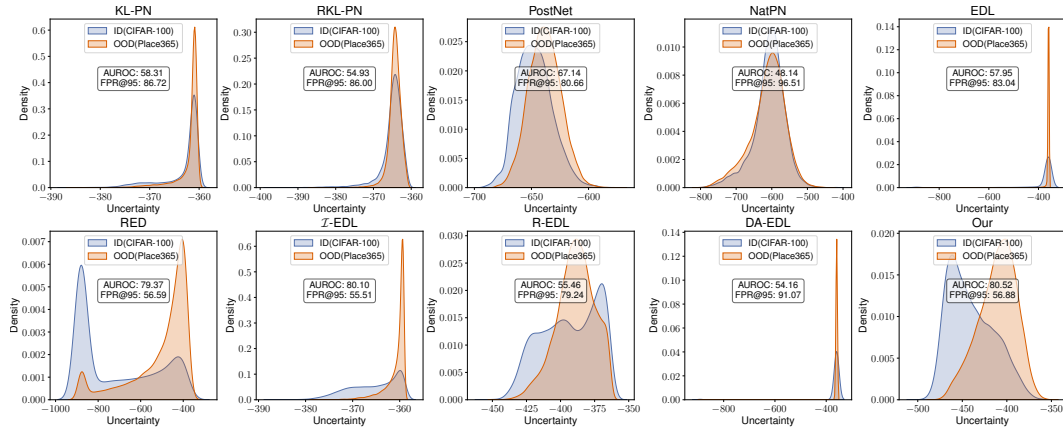


Figure 21: Differential entropy distribution on CIFAR100 (ID) vs Place365 (OOD).

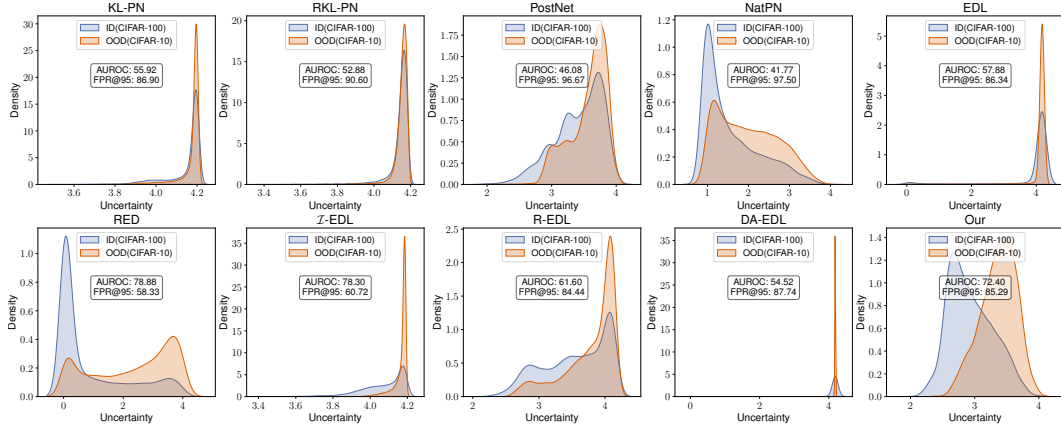


Figure 22: Mutual information distribution on CIFAR100 (ID) vs CIFAR-10 (OOD).

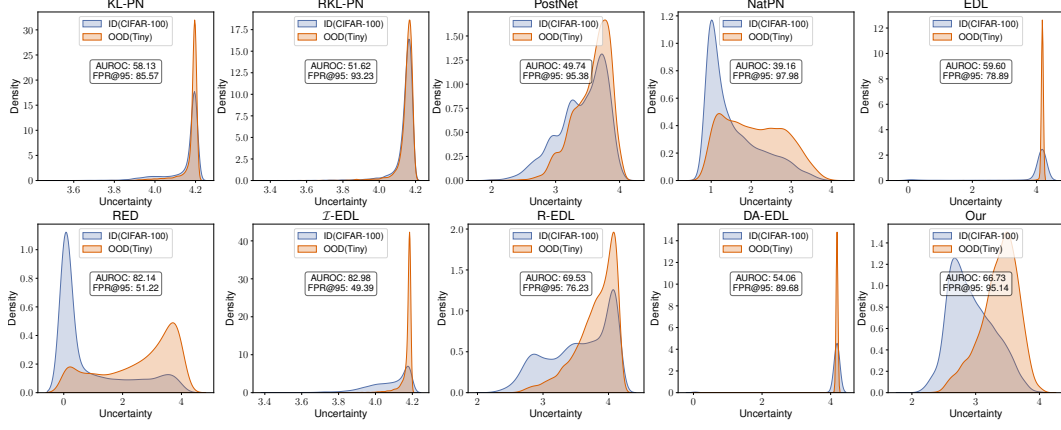


Figure 23: Mutual information distribution on CIFAR100 (ID) vs Tiny-ImageNet (OOD).

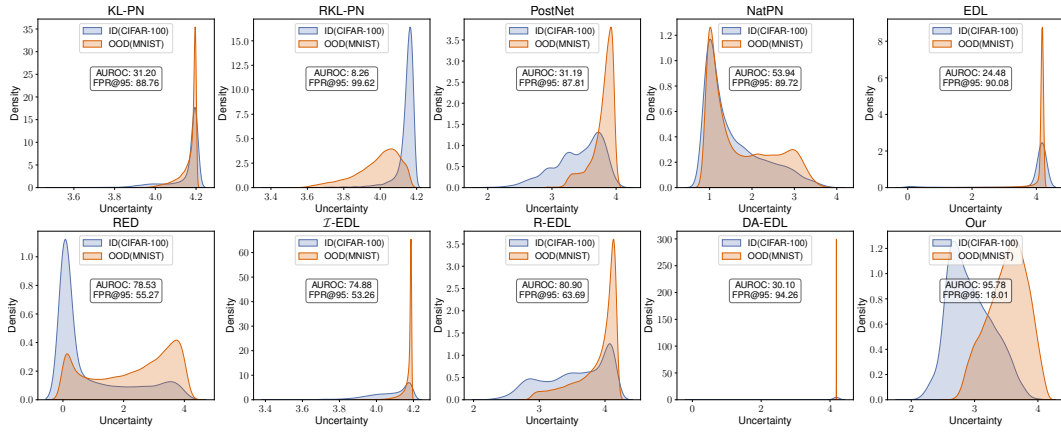


Figure 24: Mutual information distribution on CIFAR100 (ID) vs MNIST (OOD).

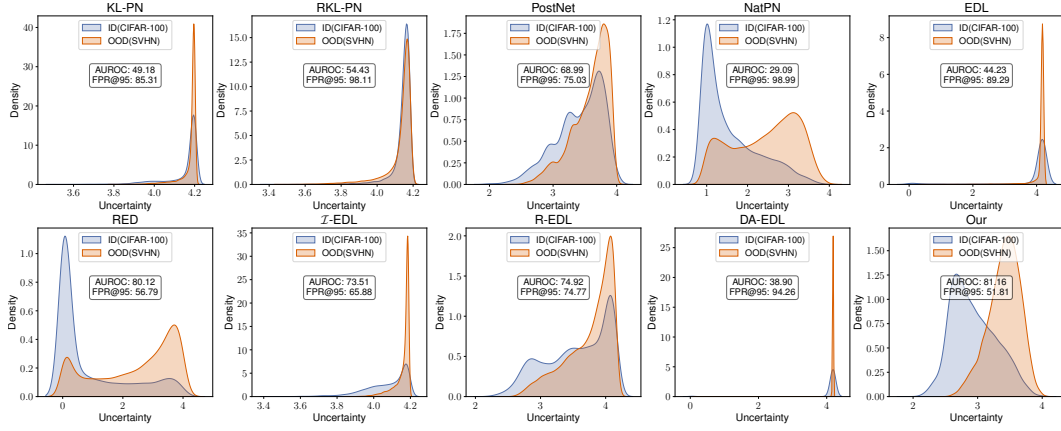


Figure 25: Mutual information distribution on CIFAR100 (ID) vs SVHN (OOD).

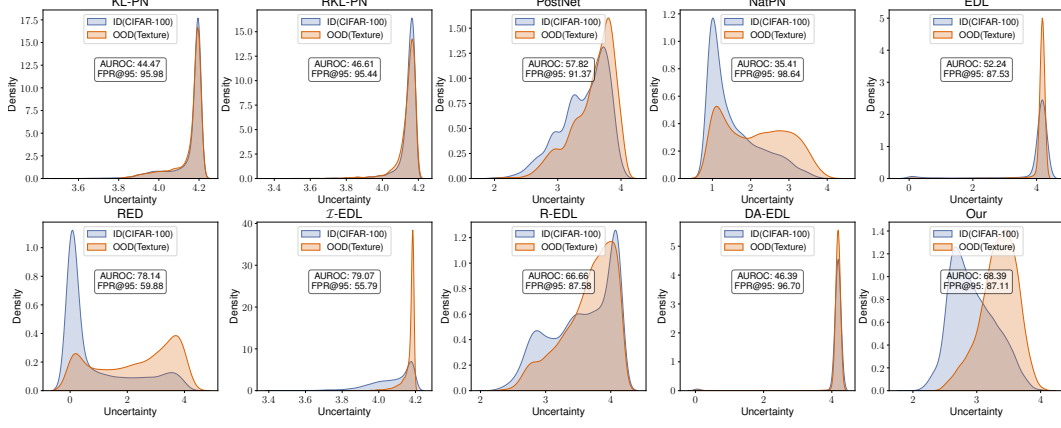


Figure 26: Mutual information distribution on CIFAR100 (ID) vs Texture (OOD).

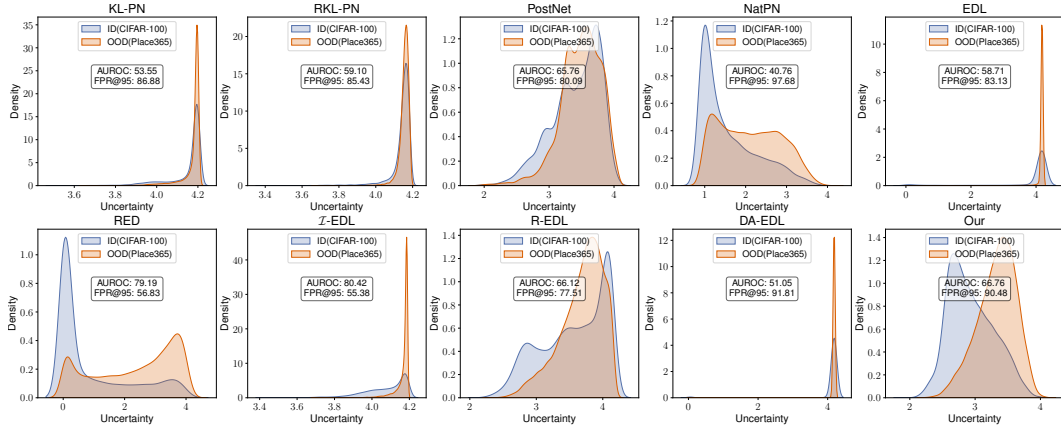


Figure 27: Mutual information distribution on CIFAR100 (ID) vs Place365 (OOD).