Learning Rate Informed Priors for Neural Network Calibration

Mouhamadou Makhtar Fall[1], Samia Ainouz[1], Pierre-Jean Lapray[2], and Jean-Philippe Tarel[3]

[1] LITIS Laboratory, INSA Rouen Normandie, Univ Rouen Normandie, Normandie Univ, F-76000 Rouen, France
[2] Université de Haute-Alsace, IRIMAS UR 7499, Mulhouse, F-68100, France
[3] COSYS/PICS-L, Univ. Gustave Eiffel, F-77454 Marne-la-Vallée, France

Abstract. Larger learning rates are frequently associated with better generalization, as they encourage exploration of flatter minima in the loss landscape. In this work, we show that the large learning rate phase produces predictive distributions that naturally capture dataset-dependent uncertainty and structured relationships between classes. Across several datasets, we observe that models trained with a large learning rate exhibit significantly lower calibration error early in training, indicating that large learning rate dynamics generate probability distributions that are more faithful to the intrinsic uncertainty of the data labels. Building on this observation, we introduce Learning Rate Informed Priors (LRIP), a simple training strategy that reuses the predictive distributions obtained at the end of the large learning rate phase as improved supervision targets. LRIP replaces uniform or heuristic smoothing with sample-dependent priors extracted directly from the model when its predictions are the most calibrated. Extensive experiments across different modalities and learning-rate schedules demonstrate that LRIP consistently improves calibration while maintaining competitive accuracy. These findings highlight a new perspective on the role of large learning rates: beyond optimization benefits, they induce informative predictive uncertainty that can be directly leveraged to build more reliable and better-calibrated models.

1 Introduction

The learning rate (LR) plays a central role in deep neural network training. Its value affects optimization stability, the scale of parameter updates, and the type of solutions reached at convergence. A large learning rate typically refers to a regime in which the LR is set close to the maximum value that still allows stable training, often several orders of magnitude larger than the values used during the refinement phase. Such regimes have been extensively studied, notably in the context of super-convergence [24], where high learning rates induce rapid exploration of the loss landscape and can lead to faster or improved generalization. Consequently, most existing work has focused on understanding these optimization dynamics, emphasizing mechanisms such as escaping sharp

minima, increased gradient noise, or transitions in dynamical regimes [10, 14, 9]. However, these perspectives primarily describe how parameters evolve, and do not address what kind of information is encoded in the model's predictions during training.

In supervised learning, predicted probabilities are as important as the final predicted label. Their quality is assessed through *calibration*, which measures how well predicted confidences reflect true likelihoods: a well-calibrated model reporting 90% confidence should be correct approximately 90% of the time [7]. Although calibration is increasingly recognized as a key property for reliable deep models, its behavior during training remains only partially understood. A key factor behind miscalibration lies in the training objective itself: standard cross-entropy optimization encourages predicted probabilities to match one-hot targets, implicitly pushing the model toward increasingly confident predictions. While this improves classification accuracy, it often leads to overconfident outputs that fail to properly reflect uncertainty, especially in later stages of training. Existing remedies either apply post-hoc corrections such as temperature scaling, or inject hand-crafted uncertainty through label smoothing and its variants [25, 16, 21]. Both families rely on externally defined targets that do not depend on the actual data distribution encountered during training.

In this work, we take a different perspective and focus on the predictive distributions produced during training itself. Through a systematic empirical analysis across multiple datasets and tasks, we observe a consistent and previously underexplored phenomenon: during the early stages of training when the learning rate is large, neural networks produce significantly better-calibrated predictions than at later stages. In this regime, the model assigns non-negligible probabilities to semantically or visually related classes, resulting in softer and more structured predictive distributions. These distributions naturally capture dataset-dependent ambiguities and encode relationships between classes that are not present in one-hot labels or uniformly smoothed targets [25]. However, as training progresses and the learning rate decreases, this property gradually disappears, leading to increasingly overconfident predictions.

We introduce Learning Rate Informed Priors (LRIP), a simple training strategy that reuses these early predictions as sample-specific supervision targets during later stages of training. By replacing fixed priors with data-dependent ones, LRIP improves the modeling of class uncertainty while preserving a lightweight and practical training procedure.

Our contributions are as follows:

1. **Characterization of the large LR phase**: we show that early predictions obtained under a large learning rate capture dataset-specific ambiguity and exhibit significantly better calibration than predictions produced later in training.
2. **A principled training strategy**: we propose LRIP, which reuses the predictive distributions at the large LR stage as informative priors for subsequent training via a KL regularization term, leading to improved supervision compared to one-hot or uniformly smoothed labels [25].

3. **Evaluation across modalities**: we demonstrate that LRIP consistently improves calibration on image classification, text classification, and semantic segmentation tasks, comparing favorably to recent calibration-aware methods such as MbLS [16], ACLS [21], and CALS [17].

This work highlights that the large learning-rate phase is not only an optimization mechanism, but also a source of informative predictive distributions that can be exploited to enhance model calibration.

2 Related work

Large learning rates in Neural Network training
Large learning rates (LRs) have been widely studied for their impact on optimization behaviour and generalization. Several works analyze how large LR values influence the training trajectory and the types of solutions reached by stochastic gradient descent (SGD) [5, 1], including studies on learning dynamics [13] and learning order effects [14]. Practical schedules such as cyclical LRs [23], warm restarts [18], and large batch scaling rules [6] further highlight the importance of the LR regime for effective training. Our work differs in focus: rather than studying optimization geometry or dynamics, we examine the predictive distributions produced during the large LR phase and how they can be reused.
Calibration-oriented training approaches
Model calibration has been addressed through post-hoc techniques like Temperature Scaling [22, 7, 3] and through training-time regularizers such as Label Smoothing [25], Focal Loss calibration [19], and adaptive or class-conditional formulations including MbLS [16], CALS [17], and ACLS [21]. These methods alter the target distributions or gradients to encourage better confidence estimates during training to improve calibration.
Positioning of our contribution
LRIP is complementary to these methods. Instead of designing new smoothing rules or confidence penalties, LRIP reuses the model's own predictions obtained during the large LR phase as sample-specific priors. To our knowledge, the idea of exploiting early large LR predictions as informative supervisory targets has not been explored in prior calibration or large LR studies.

3 Training loss and calibration background

3.1 Cross-Entropy loss and model calibration

Given an input $\mathbf{x}$ and its corresponding ground-truth label y, a deep neural network (DNN) outputs a logit vector $\mathbf{z} \in \mathbb{R}^C$, where C denotes the number of classes. The predicted probability for class j, also called label confidence, is obtained through the softmax function:

$$p_j = \frac{\exp(z_j)}{\sum_{k=1}^{C} \exp(z_k)}, \tag{1}$$

Training is typically performed using the cross-entropy (CE) loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k=1}^{C} q_k \log p_k, \tag{2}$$

where q is the target distribution (usually one-hot). The gradient with respect to the logit vector element z_j is:

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial z_j} = p_j - q_j. \tag{3}$$

The CE loss encourages the model to minimize the discrepancy between the predicted probability distribution $\mathbf{p}$ and the target distribution $\mathbf{q}$, thereby facilitating the learning of discriminative feature representations. Ideally, the probability p_j corresponding to the true class j should approach one, matching the target q_j. To achieve this, the CE loss drives an increase in the corresponding logit vector element z_j (and consequently in the probability p_j).

Model Calibration and Metrics

Model calibration refers to the agreement between a classifier's predicted confidence and its empirical correctness. A classifier is said to be well-calibrated when predictions made with a given confidence level are correct with approximately the same frequency. A standard quantitative formulation of this notion is provided by the Expected Calibration Error (ECE) [19], which measures the discrepancy between predicted confidence and observed accuracy.
Consider a set of N predictions, each associated with a predicted label $\hat{y}_i$ and an associated confidence $\hat{p}_i \in [0, 1]$. To estimate calibration, predictions are partitioned into M disjoint confidence bins $\{B_m\}_{m=1}^{M}$, each corresponding to a fixed confidence interval.
For a given bin B_m, the empirical accuracy is defined as

$$A_m = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i),$$

while the average confidence is given by

$$C_m = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.
The Expected Calibration Error is then defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |A_m - C_m|. \tag{4}$$

Lower values of ECE indicate better calibration. Over-confident models correspond to bins where $C_m > A_m$, while under-confident models satisfy $C_m < A_m$. A closely related variant, Adaptive ECE (AECE)[20], follows the same formulation but uses bins constructed to contain approximately equal numbers of samples, providing a complementary view of calibration behaviour.

Illustrative example (medical thresholding). Consider an automated system for detecting malignant lesions in medical images. Clinicians choose to automatically accept predictions with confidence $\hat{p} \geq 0.90$, and send the rest for manual review.

- **If the model is well-calibrated (low ECE/AECE):** Predictions above 0.90 confidence are truly correct about 90% of the time, meaning that among 100 such predictions the model makes about 90 correct decisions and 10 errors, in average.
- **If the model is over-confident (high ECE/AECE):** The model still reports confidence above 0.90, but its actual accuracy in this range drops in fact to only 60%. Thus, for the same 100 high-confidence predictions, it now produces 60 correct decisions and 40 errors, in average.

This example shows that accuracy alone is insufficient in high-stakes settings: it is the calibration and thus the reliability of the confidence scores that determines whether threshold-based decisions are safe or not.

4 Proposed Method

The learning rate (LR) strongly influences the dynamics of neural network training. Large learning rates induce an exploratory regime in which model predictions remain relatively soft, assigning non-negligible probabilities to multiple classes. Such predictive distributions can be interpreted as encoding dataset-dependent uncertainty and class relationships.

This perspective establishes a direct connection with calibration-oriented training methods. Approaches such as Label Smoothing [25] introduce uncertainty through predefined priors, typically uniform, which are agnostic to the data distribution. As a result, they fail to capture sample-specific ambiguities.

We instead propose to leverage the model's own predictive distributions obtained during the large learning rate phase as data-dependent priors. In the following, we formalize this idea and introduce Learning Rate Informed Priors (LRIP), a simple training strategy that integrates these distributions into the learning objective. Label smoothing [25] and its adaptive variants attempt to compensate for the limitations of one-hot labels by injecting uncertainty manually. In contrast, the early predictions obtained under a large learning rate arise naturally from the learning capturing relevant a class ambiguities. Our goal in this section is to propose how these predictions can be integrated into the learning objective.

4.1 Label Smoothing as a Probabilistic Objective

Given a one-hot target vector $y \in \mathbb{R}^K$, a prediction vector $p \in \mathbb{R}^K$, and a smoothing factor $\varepsilon \in [0, 1[$, Label Smoothing (LS) [25] replaces the one-hot hard target with a softened distribution. Here, K denotes the number of classes and the index $j \in \{1, \ldots, K\}$ refers to a specific class. In LS, the smoothed target is defined as:

$$y_j^{LS} = (1 - \varepsilon)\, y_j + \frac{\varepsilon}{K}.$$

The corresponding loss is now:

$$\mathcal{L}_{LS} = -\sum_{j=1}^{K} y_j^{LS} \log p_j,$$

where p_j denotes the predicted class probability defined in Eq. (1).
By expanding this expression and using the definition of the cross-entropy loss in Eq. (2) together with the Kullback–Leibler divergence, the LS objective can be rewritten as:

$$\mathcal{L}_{LS} = (1 - \varepsilon)\, \mathcal{L}_{CE}(y, p) + \varepsilon\, D_{\mathrm{KL}}(u_{\mathrm{uni}} \,\|\, p) + \varepsilon \log K,$$

where $u_{\mathrm{uni}} = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$ is the uniform distribution over the K classes and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence.
This formulation makes explicit that LS combines the standard cross-entropy objective with a KL regularization term toward a fixed prior, yielding a probabilistic interpretation of label smoothing.

4.2 Learning Rate Informed Priors (LRIP)

We propose to replace the fixed uniform prior with a data-dependent prior extracted from the model during training. LRIP relies on a simple observation: learning-rate schedules typically include an exploratory phase where the LR is large (e.g., warm-up or OneCycle policies), followed by a refinement phase where the LR decreases. We define $t^\star$ as the iteration at which the learning rate reaches its maximum.

At this point, we extract the predictive distribution for each training sample:

$$u_{\mathrm{LR}}(x_n) = p(x_n; \theta_{t^\star}),$$

where $\theta_{t^\star}$ are the model parameters at the LR peak. These distributions are then used as sample-specific priors in the training objective:

$$\mathcal{L}(x_n) = (1 - \varepsilon)\, \mathcal{L}_{CE}(y_n, p_n) + \varepsilon\, D_{\mathrm{KL}}(u_{\mathrm{LR}}(x_n) \,\|\, p_n)\,.$$

This formulation preserves the discriminative objective while encouraging consistency with the uncertainty captured during the exploratory phase.

4.3 Training Procedure

Training proceeds in two operational stages (see Fig. 1 for an overview):

Exploratory stage (large learning rate). The model is trained using the standard cross-entropy loss while the learning rate remains at or increases toward its maximum value. This phase which depends only on the schedule continues until the learning rate reaches its peak. At this point, indexed by $t^\star$, we record the extracted predicted confidence $u_{\mathrm{LR}}(x_n)$ for every training sample. The proposed rule requires no additional hyperparameters beyond the smoothing strength ε used in the final objective.

Refinement stage (post-peak). When the learning rate begins to decrease, training switches to the LRIP objective. The KL term regularizes the model toward the stored priors, allowing the network to refine decision boundaries while being aware of uncertainties obtained during the first phase exploration.

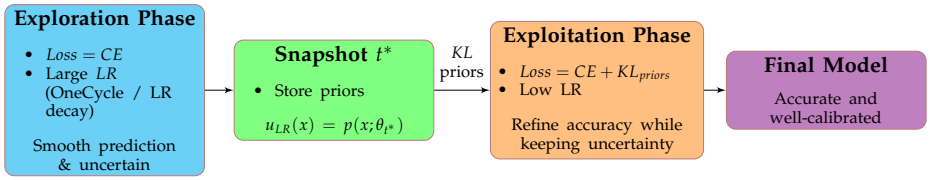


Fig. 1: Overview of the LRIP method. Early predictions obtained at the large learning rate phase are used as sample-specific priors during the refinement stage.

4.4 Computational Cost and Implementation Details

LRIP introduces minimal overhead compared to standard training.

Additional computation. The method requires a single additional forward pass over the training set at $t^\star$ to extract predictive distributions. This operation has a cost equivalent to one evaluation epoch and does not involve backpropagation.

Storage strategy. In our implementation, the predictive distributions $u_{\mathrm{LR}}(x_n)$ are computed once at $t^\star$ and stored for all training samples. These stored priors are then reused during the refinement phase without recomputation. This incurs a one-time memory cost proportional to $\mathcal{O}(N \times K)$, where N is the number of samples and K the number of classes, which remains constant throughout training.

Overall cost. LRIP does not introduce additional trainable parameters, does not require hyperparameter tuning beyond ε, and adds only minimal computational overhead corresponding to a single forward pass over the dataset.

5 Experiments

We structure our experimental evaluation in two parts. We first conduct a controlled study to empirically motivate the design of LRIP by isolating the effect of learning rate magnitude on calibration (Section 5.2). We then evaluate LRIP against existing calibration methods across multiple datasets and modalities (Section 5.3).

5.1 Experimental Setup

Datasets. We conduct experiments on CIFAR-10 [11], Tiny-ImageNet [**?**], CUB-200-2011 [26], 20 Newsgroups [12], and PASCAL VOC 2012 [4]. CIFAR-10 contains 50,000 training and 10,000 test images across 10 classes. Tiny-ImageNet includes 100,000 training images and 10,000 test images over 200 classes. CUB-200-2011 consists of 11,788 fine-grained bird images across 200 categories. The 20 Newsgroups dataset contains 20,000 text documents from 20 topics, several of which are semantically close. PASCAL VOC 2012 is a standard semantic segmentation benchmark with 20 object categories.

Architectures and optimizers. For CIFAR-10 and Tiny-ImageNet, we use ResNet-50 [8]; for CUB-200-2011, ResNet-101 [8]; for 20 Newsgroups, a GPCN architecture [15] with pretrained GloVe embeddings; for PASCAL VOC 2012, DeepLabV3 [2] with an ImageNet-pretrained ResNet-50 backbone. Batch sizes are set to 128 for CIFAR-10, 64 for Tiny-ImageNet, 16 for CUB-200-2011, 64 for 20 Newsgroups, and 8 for PASCAL VOC 2012. Optimization uses SGD (momentum 0.9, weight decay 5×10^{-4}) for vision tasks and Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) for 20 Newsgroups, following standard choices in the calibration literature [16, 21].

Metrics. For all datasets, we report Top-1 Accuracy (ACC), Expected Calibration Error (ECE, %), and Adaptive ECE (AECE, %). For semantic segmentation, we additionally report mean Intersection-over-Union (mIoU, %). Lower ECE and AECE indicate better calibration.

LRIP hyperparameters. The smoothing strength is fixed to $\varepsilon = 0.1$ for all datasets except 20 Newsgroups where $\varepsilon = 0.3$. The snapshot epoch $t^\star$ is fully determined by the learning-rate schedule and requires no additional tuning: for step-decay schedules, $t^\star$ corresponds to the first decay epoch; for OneCycle, $t^\star = \texttt{pct_start} \times T$. Concretely: $t^\star = 50$ for CIFAR-10 (150 epochs, decay at 50 and 100), $t^\star = 40$ for Tiny-ImageNet (100 epochs, decay at 40 and 60), $t^\star = 48$ for CUB-200-2011 (OneCycle, 160 epochs, pct_start=0.3), $t^\star = 9$ for 20 Newsgroups (OneCycle, 30 epochs, pct_start=0.3), and $t^\star = 40$ for PASCAL VOC 2012 (100 epochs, decay at 40 and 80). For image and text classification, each model is trained with three different random seeds and results are averaged. For PASCAL VOC 2012, we directly reuse the results reported by MbLS [16] and ACLS [21] under their original protocol.

5.2 Effect of Learning Rate Magnitude on Calibration

Before evaluating LRIP against competing methods, we isolate the effect of learning rate magnitude on calibration through a controlled study using fixed learning rates throughout training. This removes any confound from scheduling and directly exposes the relationship between LR magnitude and the quality of predictive distributions.

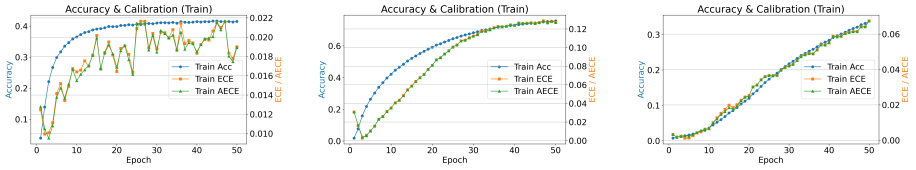
We study three LR values per dataset representing large, intermediate, and small ranges, using the same architectures and optimizers described above, trained for 50 epochs with a constant LR.

Results. Table 1 and Fig. 2 show that LR magnitude directly influences calibration behaviour. Large learning rates produce lower ECE and AECE values even when accuracy is still increasing: predicted confidences remain soft and spread across related classes, reflecting dataset ambiguity. Intermediate LRs yield the highest accuracy but consistently lead to higher calibration error, as outputs sharpen early and become overconfident. Very small LRs underfit both accuracy and calibration without producing meaningful probability structures. This pattern is consistent across all four datasets.

Table 1: Effect of a fixed learning rate on training accuracy and calibration. Best values per dataset are **bolded**.

Dataset	LR	ACC_train	ECE_train	AECE_train
CIFAR-10	0.1	85.34	0.70	0.58
	0.01	**94.52**	**0.17**	**0.12**
	0.001	70.92	0.68	0.67
Tiny-ImageNet	0.1	41.55	**1.92**	**1.89**
	0.01	**74.09**	12.56	12.56
	0.001	33.24	6.49	6.49
CUB-200-2011	0.1	20.13	**3.32**	**3.83**
	0.01	**57.32**	11.50	11.50
	0.001	0.87	0.15	0.32
20 Newsgroups	0.01	65.79	1.60	1.60
	0.001	**95.65**	**0.62**	**0.69**
	0.0001	79.25	6.47	6.45

Discussion. Large LRs generate predicted confidences that remain spread over related classes, better reflecting dataset ambiguities and leading to substantially lower calibration error. Intermediate and small LRs either collapse too quickly into overconfident predictions or fail to explore the data effectively. These observations directly motivate LRIP: if predictions under a large LR encode meaningful uncertainty and class relationships, capturing them at $t^\star$ provides informative priors for the refinement phase.



(a) ECE evolution (LR = 0.1) (b) ECE evolution (LR = 0.01) (c) ECE evolution (LR = 0.001)

Fig. 2: Effect of a fixed learning rate on Tiny-ImageNet calibration values. Each plot shows how ECE varies throughout training, highlighting the distinct behaviour induced by large, intermediate, and small LR values.

5.3 Comparison with Calibration Methods

We now evaluate LRIP against cross-entropy (CE), Label Smoothing (LS) [25], MbLS [16], ACLS [21], and CALS [17] across all datasets.

Image Classification. Table 2 summarizes results on CIFAR-10, Tiny-ImageNet, and CUB-200-2011.

CIFAR-10. LRIP provides the lowest calibration errors among all methods. For ResNet-50, LRIP achieves an ECE of **1.25** compared to 3.35 (CE), 3.66 (LS), and 1.44 (MbLS). For ResNet-101, LRIP again obtains the best ECE of **1.22**, outperforming CE (3.37), LS (3.43), and MbLS (1.58). Accuracy remains comparable to all baselines, indicating that calibration improvements do not come at the expense of discriminative performance.

Tiny-ImageNet. This dataset contains more fine-grained ambiguity, yet LRIP continues to provide competitive calibration. With ResNet-50, LRIP achieves ECE = **1.55**, second only to ACLS (1.35). With ResNet-101, LRIP reaches ECE = 1.53, close to the best-performing ACLS (1.21) and significantly better than CE (4.97) or LS (2.14). These results show that priors extracted under large LRs remain informative even in high-entropy settings.

CUB-200-2011. CUB is a fine-grained dataset with strong inter-class similarity. LRIP achieves the best calibration among all methods with ECE = **1.77**, outperforming CE (9.53), LS (3.89), MbLS (2.10), and ACLS (2.81). Although accuracy is slightly lower (67.27% vs. 68.97% for CE), this trade-off is expected: on a dataset with 200 visually similar classes, LRIP regularizes more strongly toward soft class distributions, which substantially reduces overconfidence at the cost of a modest accuracy reduction. This behaviour aligns with the known tension between discriminative sharpness and calibration in fine-grained recognition. CALS is omitted as it requires a validation set unavailable in this protocol.

Table 2: Image tasks: Results on CIFAR-10, Tiny-ImageNet and CUB-200-2011.

Method	CIFAR-10 (R50)			CIFAR-10 (R101)			Tiny-ImageNet (R50)			Tiny-ImageNet (R101)			CUB-200-2011 (R101)		
	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓
CE	94.65	3.35	3.35	94.73	3.37	3.36	64.77	3.37	3.35	65.62	4.97	4.36	68.97	9.53	9.53
LS [25]	94.66	3.66	4.00	94.76	3.43	4.12	64.81	2.22	2.27	65.97	2.14	2.18	70.22	3.89	3.90
MbLS [16]	94.45	1.44	3.42	94.75	1.58	3.36	64.69	2.19	2.30	65.69	1.39	1.18	69.72	2.10	2.14
ACLS [21]	94.68	2.26	3.02	94.57	2.53	3.31	64.83	**1.35**	**1.31**	65.72	**1.21**	**1.04**	69.84	2.81	2.64
CALS [17]	94.17	2.34	3.10	94.20	2.65	3.23	65.12	4.11	3.98	66.18	3.91	3.79	-	-	-
LRIP	94.24	**1.25**	**1.21**	94.51	**1.22**	**1.24**	64.30	1.55	1.55	65.25	1.53	1.42	67.27	**1.77**	**2.00**

Text Classification. Table 3 reports results on 20 Newsgroups. LRIP clearly outperforms all competing approaches in calibration, reducing AECE to **2.99** compared to 16.32 (CE), 4.75 (LS), 4.83 (MbLS), and 8.16 (ACLS). Accuracy remains effectively identical across all methods, demonstrating that the calibration improvement stems from a better uncertainty snapshot at $t^\star$ rather than from changes in predictive performance.

Table 3: Results on the testing set of the 20 Newsgroups dataset.

	CE	LS [25]	MbLS [16]	ACLS [21]	**LRIP**
ACC ↑	69.61	70.69	71.28	71.31	70.67
AECE ↓	16.32	4.75	4.83	8.16	**2.99**

Image Segmentation. On PASCAL VOC 2012 (Table 4), LRIP reduces ECE from 8.26 (CE), 9.35 (LS), 7.94 (MbLS), and 7.29 (ACLS) down to **6.59**, and AECE to **6.70**. Segmentation performance is preserved with mIoU of 71.17, confirming that calibration gains do not come at the expense of task accuracy.

Table 4: Results on PASCAL VOC 2012 segmentation benchmark.

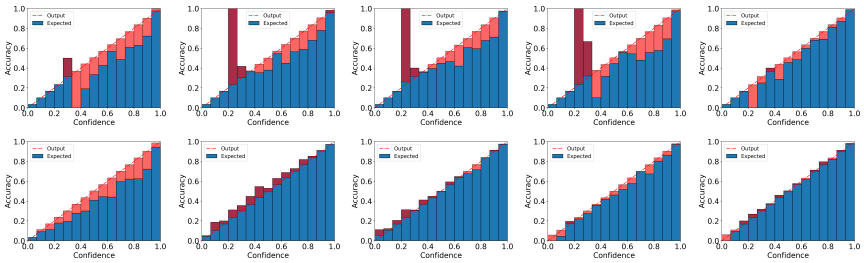
	CE	LS [25]	MbLS [16]	ACLS [21]	**LRIP**
mIoU ↑	70.92	71.00	71.20	70.63	**71.17**
ECE ↓	8.26	9.35	7.94	7.29	**6.59**
AECE ↓	8.23	9.95	7.99	7.28	**6.70**

Summary. Across all datasets and architectures, LRIP consistently yields lower ECE and AECE than CE and LS, and often surpasses more complex adaptive

smoothing methods such as MbLS or ACLS. Unlike these methods, which rely on margin-based formulations sensitive to dataset characteristics and model architecture, LRIP avoids explicit margin design and maintains stable calibration improvements across vision and language benchmarks.

5.4 Discussion

Reliability diagrams. The reliability diagrams in Fig. 3 illustrate clear differences in calibration behaviour across methods. On both CIFAR-10 (top row) and CUB-200-2011 (bottom row), CE exhibits strong over-confidence: predicted confidence consistently exceeds empirical accuracy. LS reduces this effect but remains imperfect in low- and mid-confidence bins. MbLS and ACLS partially correct these deviations but show irregular bin-level fluctuations, reflecting their sensitivity to class imbalance and dataset granularity. In contrast, LRIP produces diagrams that remain close to the ideal diagonal across all bins, confirming that the priors extracted during the large LR phase impose a coherent probabilistic structure that persists throughout training.



(a) CE (b) LS (c) MbLS (d) ACLS (e) LRIP

Fig. 3: Reliability diagrams for CIFAR-10 (top) and CUB-200-2011 (bottom).

Ablation Study: Importance of the large LR Snapshot. To confirm that the effectiveness of LRIP depends specifically on the large-LR phase, we compare it against a delayed variant **SD** that extracts priors at the second learning-rate decay (epoch 100 for CIFAR-10, epoch 60 for Tiny-ImageNet), where the model has become substantially more confident.

Table 5 shows that on CIFAR-10, LRIP achieves ECE = **1.25** while SD more than doubles it to 2.67 (AECE: 1.21 vs. 2.63). On Tiny-ImageNet, LRIP gives ECE = **1.55** vs. 2.97 for SD. In both cases, SD slightly increases accuracy, confirming that the degradation is not discriminative but reflects the collapse of late-stage predictions toward low-entropy distributions. This demonstrates that the timing of the snapshot is essential: only the large LR phase provides the calibrated uncertainty estimates needed for effective regularization.

Table 5: Ablation Study: LRIP vs. SD (Second-Decay snapshot). Late-stage priors significantly worsen calibration.

	CIFAR-10 (R50)		**Tiny-ImageNet (R50)**	
	LRIP	**SD**	**LRIP**	**SD**
ACC ↑	94.24	94.55	64.30	65.01
ECE ↓	**1.25**	2.67	**1.55**	2.97
AECE ↓	**1.21**	2.63	**1.55**	2.86

6 Conclusion

In this work, we showed that the predicted confidences obtained during the large learning rate phase contain rich and reliable information about dataset class uncertainty. By capturing these early predictions and using them as data-dependent priors, LRIP provides a simple yet principled alternative to classical calibration techniques based on fixed or hand-crafted priors. Across image classification, text classification, and semantic segmentation, LRIP consistently improves calibration without sacrificing accuracy and without introducing additional hyperparameters beyond the smoothing strength also used in Label Smoothing.
Beyond its empirical performance, this work highlights a broader and largely unexplored perspective: large learning rates are not merely an optimization tool but also a source of meaningful information about class uncertainties. Leveraging this information opens a new direction for calibration research, in which uncertainties are extracted during training itself rather than imposed externally. An important direction for future work is to investigate the robustness of LRIP under distributional shifts, including out-of-distribution and domain-shift scenarios. Since LRIP relies on uncertainty patterns learned during early training rather than fixed priors or explicit margins, it may offer improved resilience when the test distribution deviates from the training data. Exploring this behavior in safety-critical and real-world deployment settings remains a promising avenue for further research, in our opinion.

Acknowledgements This work is supported by the French National Research Agency (ANR) project INARI- ANR-23-CE22-0002. Part of this work was performed using computing resources of CRIANN (Normandy, France).

References

1. Bottou, L.: Stochastic gradient descent tricks. In: Neural networks: tricks of the trade: second edition, pp. 421–436. Springer (2012)
2. Chen, L.C.: Rethinking atrous convolution for semantic image segmentation. Computing Research Repository (2017)
3. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: Proceedings of the IEEE/CVF ICCV. pp. 6889–6899 (2021)

4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
6. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. PMLR (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE CVPR, pages=770–778, year=2016
9. Jastrzebski, S.: Three factors influencing minima in sgd. ArXiv (2017)
10. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. In: International Conference on Learning Representations (2017)
11. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
12. Lang, K.: Newsweeder: Learning to filter netnews. In: Machine learning proceedings 1995, pp. 331–339. Elsevier (1995)
13. Lewkowycz, A.e.a.: The large learning rate phase of deep learning: the catapult mechanism. arXiv preprint arXiv:2003.02218 (2020)
14. Li, Y., Wei, C., Ma, T.: Towards explaining the regularization effect of initial large learning rate in training neural networks. Advances in neural information processing systems **32** (2019)
15. Lin, M., Chen, Q., Yan, S.: Network in network. CoRR (2013)
16. Liu, B., Ben Ayed, I., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: Proceedings of the IEEE/CVF CVPR (2022)
17. Liu, B., Rony, J., Galdran, A., Dolz, J., Ben Ayed, I.: Class adaptive network calibration. In: Proceedings of the IEEE/CVF CVPR (2023)
18. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. Learning **10**, 3
19. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in neural information processing systems **33**, 15288–15299 (2020)
20. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
21. Park, H., Noh, J., Oh, Y., Baek, D., Ham, B.: Acls: Adaptive and conditional label smoothing for network calibration. In: Proceedings of the IEEE/CVF ICCV (2023)
22. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3) (1999)
23. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE WACV. IEEE (2017)
24. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE CVPR (2016)

26. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., California Institute of Technology (juill 2011)