Bayesian Deep Learning for Uncertainty Estimation in Urban Scene Perception Models

Ubaldo Taliercio
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
ubaldotaliercio282@oregonstate.edu

Alannah Mulcair
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
mulcair1996@uab.edu

Prayitna Prayoga
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
prayitnap@binghamton.edu

Abstract

Urban scene perception models now support autonomous driving, municipal mapping, accessibility auditing, and infrastructure inspection, yet deterministic segmentation and detection networks often produce overconfident outputs in occluded, rare, or distribution-shifted conditions. This manuscript develops a Bayesian deep learning framework for uncertainty estimation in dense urban scene perception, with emphasis on separating epistemic model uncertainty from aleatoric observation and annotation ambiguity. The framework treats semantic labels and object-level predictions as outcomes of a posterior predictive distribution, uses variational inference and ensemble-style posterior approximation for scalable training, and defines spatial and class-aware likelihood weighting so that small but safety-critical classes are not overwhelmed during optimisation. It further specifies a calibration layer for dense predictions, an uncertainty decomposition based on predictive entropy and mutual information, and a risk-aware decision rule that permits abstention when epistemic uncertainty is too high. The analysis clarifies why pixel-independent likelihood assumptions understate spatially correlated uncertainty, how class imbalance distorts calibration, and why post hoc temperature scaling alone cannot correct posterior inadequacy under dataset shift. The contribution is methodological and analytical rather than an empirical benchmark: it provides a coherent inference and evaluation design, identifies failure mechanisms specific to urban scenes, and links uncertainty estimates to operational decisions such as relocalisation, human review, and conservative perception fallback.

Keywords

Bayesian deep learning; uncertainty calibration; urban scene perception; semantic segmentation; distribution shift; risk-aware decision making

1. Introduction

Urban scene perception requires a model to convert high-dimensional visual and spatial observations into labels, objects, traversable regions, and other structured decisions. These outputs are consumed by systems that cannot treat all errors as equivalent: a missed pedestrian,

an overconfident boundary between road and pavement, or a misclassified construction barrier may carry very different consequences from a small error on a building facade. Deterministic neural networks trained with cross-entropy or related losses can achieve strong average accuracy while remaining poorly equipped to report when their predictions should not be trusted. Bayesian deep learning offers a principled language for this problem by placing a distribution over model parameters or predictive functions and by propagating uncertainty into downstream decisions [1]. Approximate inference through dropout [2], deep ensembles [3], and related posterior approximation families [4] has made such reasoning computationally plausible for modern vision architectures.

For dense urban prediction, however, uncertainty estimation is not a direct transfer of image classification practice. The likelihood is spatially structured, classes are strongly imbalanced, annotation ambiguity is concentrated at object boundaries and distant regions, and distribution shift can arise from weather, illumination, occlusion, construction, sensor degradation, or geographically unusual streetscape composition. Variational weight uncertainty [5] and practical variational inference for neural networks [6] provide a training foundation, while concrete dropout [7] and low-rank Gaussian approximations [8] show how approximate posteriors can be made scalable. Calibration research further shows that predictive probabilities need explicit treatment even when accuracy is high [9]. Urban sensing also increasingly combines street-level imagery with geotagged social media and other heterogeneous observations to model streetscape conditions using deep neural networks [10]. These developments motivate a focused question: how should Bayesian uncertainty be formulated, calibrated, and used in dense urban scene perception so that epistemic and aleatoric components remain interpretable and operationally useful?

This manuscript addresses that question through a methodological framework rather than a new empirical benchmark. Its central contribution is a Bayesian deep learning design for urban scene perception that couples posterior predictive inference, spatially and class-weighted training, dense calibration, uncertainty decomposition, and risk-aware decision rules. The framework is intentionally modular: it can be implemented with variational last layers, Monte Carlo dropout, deep ensembles, or a combination, provided that the resulting predictive distribution is treated as an approximation whose calibration and failure modes are explicitly audited. The analysis explains why conventional pixel-wise likelihoods can understate spatially correlated uncertainty, how class imbalance and boundary ambiguity affect calibration, and what a conservative perception system can do when epistemic uncertainty remains high.

2. Problem Formulation and Urban Scene Risk

Let x denote an urban image or image sequence, and let y denote a structured label field. In semantic segmentation, y assigns a class to each pixel or superpixel; in object detection or instance segmentation, y contains object locations, classes, and masks. A perception model parameterised by θ defines a conditional distribution p(y|x,θ). Bayesian inference replaces a single parameter estimate with a posterior distribution over parameters or predictive functions, so that the model reports a distribution over plausible scene interpretations rather than one deterministic label map. The predictive posterior for a new observation x is obtained by integrating over the posterior distribution, which makes the distinction between model uncertainty and observation noise explicit.

$$p(y*|x*,D) = \int p(y*|x*,\theta)\, p(\theta|D)\, d\theta$$

Here D denotes the training data, p(θ|D) is the posterior over parameters, and p(y|x,θ) is the observation model. The integral is generally intractable for deep networks, but it clarifies the target of any approximation. A variational approximation, dropout sample, or ensemble member should be judged by how well it represents this predictive distribution in regions that matter for decisions, not only by its average accuracy.

Urban scene uncertainty has at least two qualitatively different sources. Aleatoric uncertainty arises from sensor noise, motion blur, reflective surfaces, rain, annotation ambiguity, and the irreducible ambiguity of small or occluded objects. Epistemic uncertainty arises from limited knowledge of the model, including underrepresented geography, unusual construction layouts, rare vulnerable road users, and shift between training and deployment domains. For a scalar prediction, the total predictive variance can be decomposed into an expected conditional variance and the variance of conditional means.

$$\mathrm{Var}(y|x,D) = E_{p(\theta|D)}[\mathrm{Var}(y|x,\theta)] + \mathrm{Var}_{p(\theta|D)}[\mathrm{E}(y|x,\theta)]$$

The first term is aleatoric and the second is epistemic. For categorical scene labels, the same idea is expressed through entropy and mutual information rather than scalar variance, as discussed in Section 4. This decomposition matters because the appropriate response differs: aleatoric uncertainty may call for sensor fusion, conservative geometry, or label smoothing, whereas epistemic uncertainty may call for abstention, human review, or exploration of a new environment.

Urban datasets make these issues concrete. Cityscapes provides dense semantic annotations of complex street scenes [11]; KITTI established an early benchmark for autonomous driving perception [12]; and nuScenes introduced multimodal urban driving data with sensors and annotations designed for scene understanding [13]. These resources support the study of urban perception, but they do not eliminate domain shift. A model trained on one city, camera rig, or season can encounter different architecture, vegetation, road markings, and social practices elsewhere. The Bayesian framework must therefore be evaluated not only on in-distribution accuracy but also on the reliability of its uncertainty under plausible shift.

3. Bayesian Inference Architecture for Dense Scene Perception

A practical Bayesian urban perception model begins with a likelihood appropriate for structured outputs. For semantic segmentation, a common choice is a per-pixel categorical likelihood conditioned on logits produced by an encoder-decoder network. For detection, the likelihood may combine objectness, class, and bounding-box or mask terms. Encoder-decoder architectures such as Bayesian SegNet illustrate how model uncertainty can be propagated through dense prediction [14]. The Bayesian treatment can be applied to the full network, to the decoder, or to a last-layer stochastic head; the choice controls computational cost and the richness of the approximate posterior. Full-network variational inference is expressive but expensive, while last-layer inference is cheaper but may underrepresent representation-level uncertainty.

Let qφ(θ) denote a variational approximation with parameters φ. Training maximises a lower bound on the marginal likelihood of the data.

$$L(\varphi) = E_{q\varphi(\theta)}[\log p(D|\theta)] - KL\big(q\varphi(\theta)||p(\theta)\big)$$

The first term rewards data fit under sampled parameters, and the Kullback-Leibler term penalises departure from the prior p(θ). In dense prediction, a naive sum over pixels treats all pixel labels as independent and equally important. That assumption is analytically convenient

but poorly matched to urban scenes, where boundary pixels are ambiguous, large classes dominate the loss, and small classes such as cyclists, traffic signs, or vulnerable pedestrians carry disproportionate risk. A weighted objective can partially correct this imbalance by assigning class and location weights w_i.

$$L_w(\varphi) = \sum_{i=1}^{N} w_i E_{q\varphi(\theta)}[\log p(y_i|x_i,\theta)] - KL\big(q\varphi(\theta)||p(\theta)\big)$$

Here N is the number of labelled units, y_i is the target for unit i, x_i is the corresponding input context, and w_i encodes class rarity, boundary proximity, or decision cost. The weights should be specified before training and audited, because they change the posterior approximation itself: increasing the weight of a rare class makes the model more willing to explain that class, which can improve recall but may also increase false positives. The weighting scheme is therefore not a harmless implementation detail but a statement about which errors matter in the urban domain.

Several posterior approximation strategies are compatible with this objective. Monte Carlo dropout interprets dropout at test time as approximate Bayesian sampling [2]. Concrete dropout learns dropout rates and can reduce the burden of manual tuning [7]. Deep ensembles train multiple models with different initialisations and treat their disagreement as an uncertainty signal [3]. Low-rank or Gaussian approximations such as SWAG summarise parameter uncertainty around a trained trajectory [8]. Evidential approaches place a different output layer on the network and model uncertainty through evidence parameters [18]. These methods do not produce identical uncertainty estimates. Dropout can be sensitive to rate and placement, ensembles capture multi-modality better but multiply training and inference cost, and Gaussian approximations may struggle when the posterior is strongly non-Gaussian. The framework therefore treats the inference method as an explicit design choice to be matched to latency, memory, and safety requirements.

4. Uncertainty Decomposition and Dense Calibration

For categorical predictions, the predictive distribution is a vector of class probabilities at each pixel or object. Let p_c(x) denote the posterior predictive probability of class c at location x, and let p_c(x,θ) denote the probability under a sampled parameter setting. Epistemic uncertainty can be measured by mutual information between the label and the parameters. It equals the entropy of the mean predictive distribution minus the expected entropy of individual predictive distributions.

$$I(y;\theta|x,D) = H\big(E_{p(\theta|D)}[p(y|x,\theta)]\big) - E_{p(\theta|D)}\big[H\big(p(y|x,\theta)\big)\big]$$

The first term, H(E[p]), is total predictive entropy. The second term, E[H(p)], is expected aleatoric entropy. Their difference is a principled measure of epistemic uncertainty for classification-like outputs. In urban scenes, total entropy may be high at object boundaries because the label itself is ambiguous, while mutual information may remain low if all plausible models agree that the boundary is ambiguous. Conversely, an unusual object or a shifted roadway layout may produce high mutual information because different posterior samples disagree about the scene interpretation. This distinction is essential for triage: boundary ambiguity often warrants geometric refinement, whereas high epistemic uncertainty may warrant abstention or human inspection.

Calibration is a separate property from sharpness or accuracy. A dense model is calibrated if, among predictions assigned confidence in a given range, the observed frequency of

correctness matches that confidence. A standard expected calibration error partitions predictions into confidence bins and compares accuracy with confidence.

$$ECE = \sum_{b=1}^{B} (|S_b|/N) \, |acc(S_b) - conf(S_b)|$$

Here B is the number of bins, S_b is the set of predictions in bin b, N is the total number of predictions, acc(S_b) is the empirical accuracy in the bin, and conf(S_b) is the mean confidence. In segmentation, this calculation must be handled carefully because pixels are not independent and because large classes dominate simple averages. Class-conditioned calibration, boundary-aware calibration, and region-level calibration are more informative for urban perception than a single global score. Temperature scaling can improve calibration by dividing logits by a positive scalar learned on held-out data [9], but it cannot repair an approximate posterior that is fundamentally overconfident under distribution shift. Calibration must therefore be evaluated on shifted conditions and on rare classes, not only on the validation split drawn from the training distribution.

Posterior predictive checks provide a complementary diagnostic. If the model is asked to predict a rare class under occlusion, the predictive distribution should broaden rather than assign a sharp but incorrect label. If a scene contains an object category absent from training, the model should ideally assign high epistemic uncertainty or high novelty score rather than confidently mapping the object to a known class. Out-of-distribution detection baselines based on softmax confidence are useful reference points [15], but dataset-shift studies show that uncertainty quality can degrade substantially as shift increases [16]. The Bayesian layer should therefore be evaluated as a reliability mechanism under stress, not merely as a probability-generating add-on.

5. Inference Protocol and Evaluation Logic

The proposed protocol separates training, calibration, and decision evaluation. Training uses a weighted variational or ensemble objective to approximate the posterior over parameters. Calibration uses held-out data to fit post hoc parameters such as temperature and to estimate class-conditioned calibration curves. Decision evaluation uses a risk-coverage analysis in which the model is allowed to abstain on a fraction of predictions. For a selected set S_c that achieves coverage c, selective risk is the average loss on the retained predictions.

Here ℓ is the task loss, \hat y_i is the predicted label or object, and |S_c| is the number of retained predictions. A useful uncertainty estimator should produce a favourable risk-coverage trade-off: as coverage decreases, retained predictions should become more reliable. This analysis is more decision-relevant than a single calibration number because it reflects what a perception system can do when uncertainty is high.

The framework also requires an explicit action model. In a safety-relevant urban system, the possible actions might include accepting the label, requesting a second sensor modality, reducing speed, relocalising, or sending the scene to a human reviewer. Let A be the set of admissible actions, let C(a,y) be the cost of action a when the true label is y, and let κ be an epistemic uncertainty threshold above which the system should not make an autonomous commitment. The risk-aware decision rule selects the action that minimises expected cost among actions whose associated epistemic uncertainty is acceptable.

$$a*(x) = argmin_{a \in A_\kappa(x)} E_{p(y|x,D)}[C(a,y)], \quad A_\kappa(x) = \{a: I(y;\theta|x,D) \leq \kappa\}$$

The threshold κ is not a universal constant. It should be derived from operational risk tolerance, class-specific consequences, and the cost of deferral. A pedestrian-related decision

may require a lower threshold than a building-facade label. The cost matrix C encodes domain knowledge that accuracy metrics do not: confusing a sidewalk with road may be more consequential than confusing two vegetation classes. This decision-theoretic layer prevents uncertainty from remaining a diagnostic number and forces the model designer to state how uncertainty changes behaviour.

Evaluation should include multiple baselines: deterministic softmax confidence, Monte Carlo dropout, deep ensembles, a low-rank Gaussian approximation, temperature-scaled variants, and an evidential head where appropriate. Each should be compared on negative log-likelihood, Brier score, class-conditioned calibration error, risk-coverage behaviour, and out-of-distribution detection under controlled shifts. The comparison should report computational cost, including forward passes, memory, and training time. A method that improves uncertainty slightly but doubles latency may be inappropriate for a real-time vehicle, while a method that is cheap but poorly calibrated may be unsuitable for safety review. The protocol should also test spatial correlation: uncertainty maps that are smooth everywhere may look visually plausible yet fail to localise the specific object that caused model disagreement. Region-level and object-level analyses are therefore necessary complements to pixel-level metrics.

6. Technical Discussion: Failure Mechanisms and Design Trade-offs

The first major failure mechanism is the treatment of dense predictions as independent. Pixel-wise likelihoods and pixel-wise calibration can overstate the effective amount of evidence because neighbouring pixels share illumination, texture, and object identity. A model may become highly confident on a large homogeneous surface and still fail at its boundary or at a small object inside it. Correlated likelihoods, patch-level calibration, and object-level uncertainty aggregation are therefore important. Spatial smoothing of uncertainty can reduce noise, but excessive smoothing can erase the localized signal that identifies a missed pedestrian or a misclassified curb. The appropriate degree of spatial pooling depends on the task: lane-level decisions may benefit from region aggregation, while object detection requires localized uncertainty.

The second failure mechanism is class imbalance interacting with posterior approximation. A rare class receives few likelihood terms, so its posterior may remain broad or may collapse toward the majority class. Weighted likelihoods help, but they also alter calibration. A model can become more sensitive to rare classes while producing probabilities that are no longer well calibrated for common classes. Class-conditioned calibration and separate thresholds for rare safety-critical categories are therefore more defensible than a single global calibration correction. The discussion of probabilistic generalisation further cautions that Bayesian model evidence and predictive performance are related but not identical [19], and studies of posterior geometry show that approximate posteriors can differ substantially from the true posterior in deep networks [20]. This is not a reason to abandon Bayesian methods; it is a reason to treat them as approximate reliability tools whose behaviour must be empirically audited in the deployment domain.

The third failure mechanism is distribution shift. Rain, snow, low sun, motion blur, unusual road markings, construction sites, and geographically unfamiliar streetscape elements can all move observations away from the training distribution. Softmax confidence often remains high in such cases, and post hoc temperature scaling fitted on in-distribution validation data may not transfer. Ensemble disagreement and mutual information can provide a stronger signal, but they can also be muted if all ensemble members share the same

inductive bias or training data. Combining Bayesian uncertainty with input-space novelty detection, temporal consistency checks, and sensor-conflict monitoring is therefore more robust than relying on a single uncertainty score. The urban sensing literature illustrates that scene understanding can draw on heterogeneous observations beyond a single camera stream [10]; that heterogeneity is also an opportunity for cross-checking, because disagreement between modalities can reveal epistemic uncertainty that a unimodal model misses.

The final trade-off concerns deployment. Full variational inference over all weights is computationally demanding; Monte Carlo dropout requires multiple stochastic forward passes; deep ensembles require storing and evaluating several models. These costs can be reduced by using a Bayesian last layer, a small ensemble, or adaptive sampling in which additional forward passes are triggered only when the first-pass uncertainty exceeds a threshold. Adaptive sampling is attractive because it concentrates computation on ambiguous scenes, but it complicates latency guarantees and requires a fallback when the uncertainty estimate itself is unreliable. The design should therefore specify a maximum inference budget and a safe default action when that budget is exhausted. In urban perception, a calibrated abstention is often preferable to a confident but unsupported prediction.

7. Conclusion

This manuscript has developed a Bayesian deep learning framework for uncertainty estimation in urban scene perception models. The framework treats dense prediction as posterior predictive inference, distinguishes aleatoric from epistemic uncertainty through entropy and mutual information, and incorporates spatially and class-weighted likelihoods so that rare and safety-critical urban categories are not overwhelmed during training. It specifies calibration and risk-coverage evaluation for structured outputs, and it connects uncertainty to a decision rule that can abstain or defer when epistemic uncertainty is too high. The analysis identifies several domain-specific failure mechanisms: pixel-independent likelihoods understate spatially correlated uncertainty, class imbalance distorts calibration, post hoc scaling cannot repair posterior inadequacy under shift, and deployment constraints require an explicit inference budget and fallback policy. These points are methodological rather than empirical claims, and they define a rigorous basis for future evaluation on urban datasets and real deployment conditions. The broader implication is that uncertainty in urban scene perception should be designed as part of the perception-decision interface, not appended as an afterthought to a deterministic model.

References

1. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

2. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (PMLR Vol. 48, pp. 1050–1059).

3. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

4. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Abbass, H. A., & Nahavandi, S.

(2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297.

5. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In Proceedings of the 32nd International Conference on Machine Learning (PMLR Vol. 37, pp. 1613–1622).

6. Graves, A. (2011). Practical variational inference for neural networks. In Advances in Neural Information Processing Systems (Vol. 24, pp. 2348–2356).

7. Gal, Y., Hron, J., & Kendall, A. (2017). Concrete dropout. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3581–3590).

8. Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. In Advances in Neural Information Processing Systems (Vol. 32).

9. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR Vol. 70, pp. 1321–1330).

10. Aman, J., Matisziw, T. C., Kim, J. B., & Luo, D. (2022, April). Sensing the city: Leveraging geotagged social media posts and street view imagery to model urban streetscapes using deep neural networks. In Proceedings of the International Conference on Computer-Aided Architectural Design Research in Asia (pp. 595-604). Association for Computer-Aided Architectural Design Research in Asia.

11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3213–3223).

12. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354–3361).

13. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11621–11631).

14. Kendall, A., Badrinarayanan, V., & Cipolla, R. (2017). Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In Proceedings of the British Machine Vision Conference.

15. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations.

16. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems (Vol. 32).

17. Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., & Bennamoun, M. (2022). Hands-on Bayesian neural networks—A tutorial for deep learning users. IEEE Computational Intelligence Magazine, 17(2), 29–48.

18. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems (Vol. 31, pp. 3179–3189).

19. Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In Advances in Neural Information Processing Systems (Vol. 33).

20. Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. (2021). What are Bayesian neural network posteriors really like? In Proceedings of the 38th International Conference on Machine Learning (PMLR Vol. 139, pp. 4629–4640).