Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision

Fredrik K. Gustafsson*, Martin Danelljan‡, Thomas B. Schön*

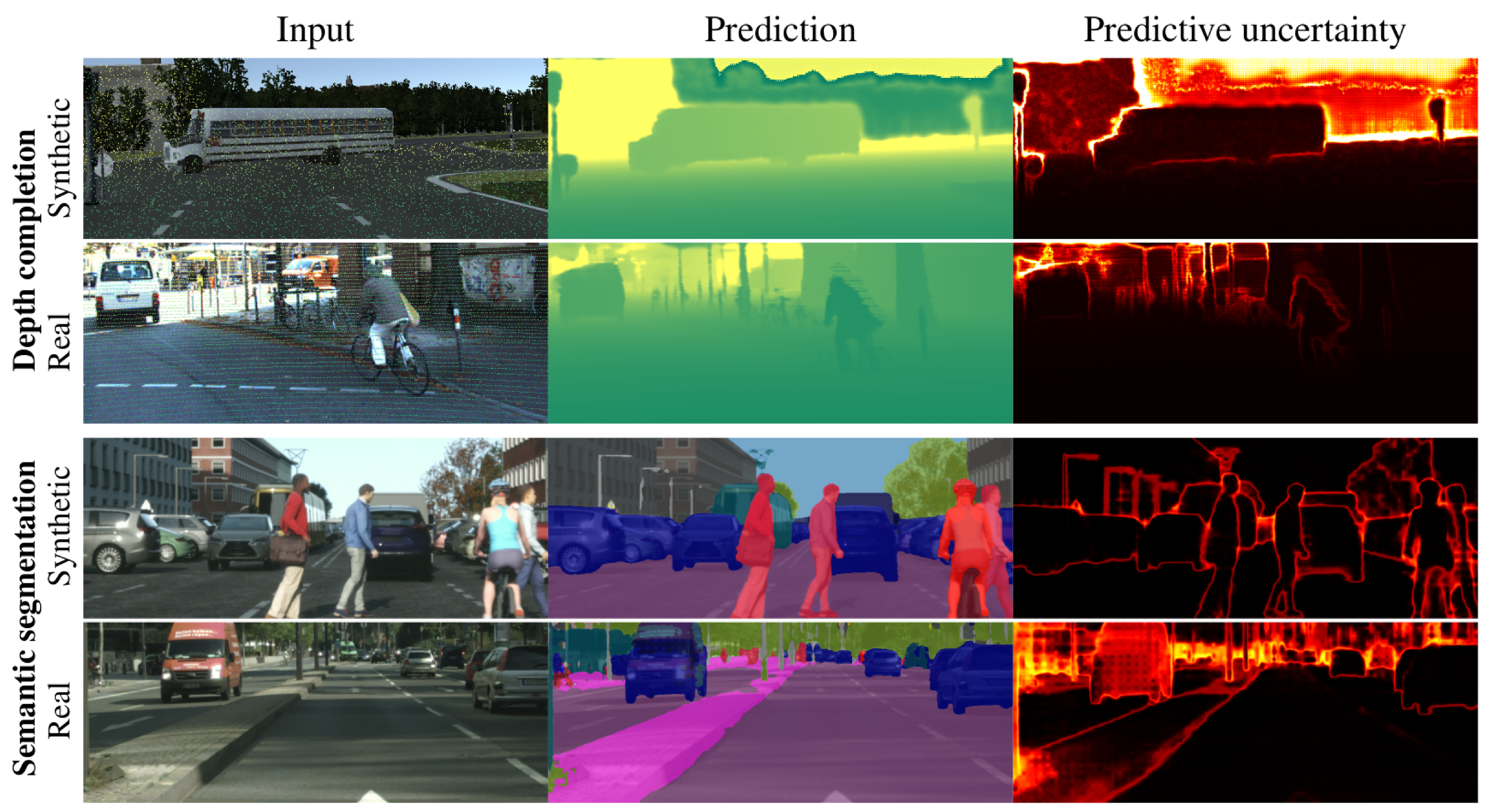
*Department of Information Technology, Uppsala University, Sweden
‡Computer Vision Laboratory, ETH Zurich, Switzerland

What is the problem?

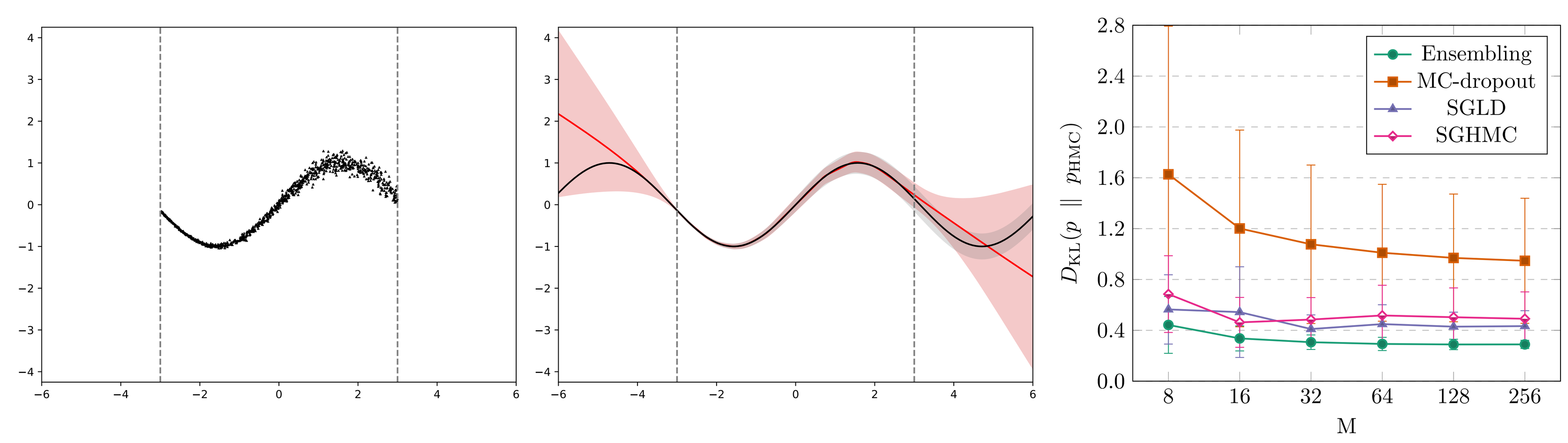
- While deep learning has become the go-to approach in computer vision, these models fail to properly capture the **uncertainty** inherent in their predictions. **Bayesian deep learning** addresses this issue in a principled manner. Predictive uncertainty is then decomposed into **aleatoric** and **epistemic** uncertainty.
- Aleatoric uncertainty captures inherent and irreducible data noise, and can be estimated by directly predicting the conditional distribution $p(y|x)$. Estimating **epistemic** uncertainty, which accounts for uncertainty in the model parameters, can mitigate model over-confidence and is thus of great importance.
- While epistemic uncertainty estimation has proven to be highly challenging, especially for *large-scale* models employed in *real-world* computer vision tasks, **scalable** techniques have recently emerged.
- The research community however lacks a common and comprehensive **evaluation framework** for such methods. Both researchers and practitioners are currently thus unable to properly assess and compare competing methods.

Our contributions

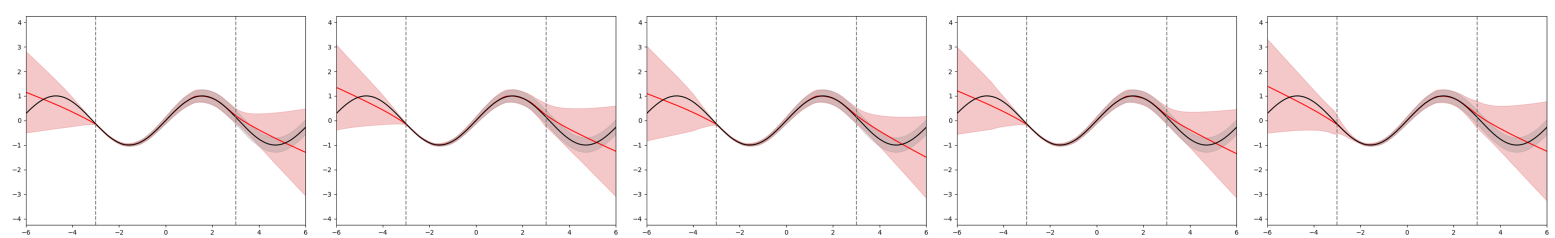
- We propose a comprehensive **evaluation framework** for *scalable* epistemic uncertainty estimation methods in deep learning. It is specifically designed to test the robustness required in **real-world** computer vision applications.
- Our proposed framework employs state-of-the-art models on the tasks of **depth completion** (regression) and **semantic segmentation** (classification).
- We provide the first properly extensive and conclusive comparison of the two current state-of-the-art *scalable* methods: **ensembling** and **MC-dropout**. Our comparison demonstrates that **ensembling** consistently provides more reliable and practically useful uncertainty estimates.
- **Publicly available source code:** `github.com/fregu856/evaluating_bdl`.



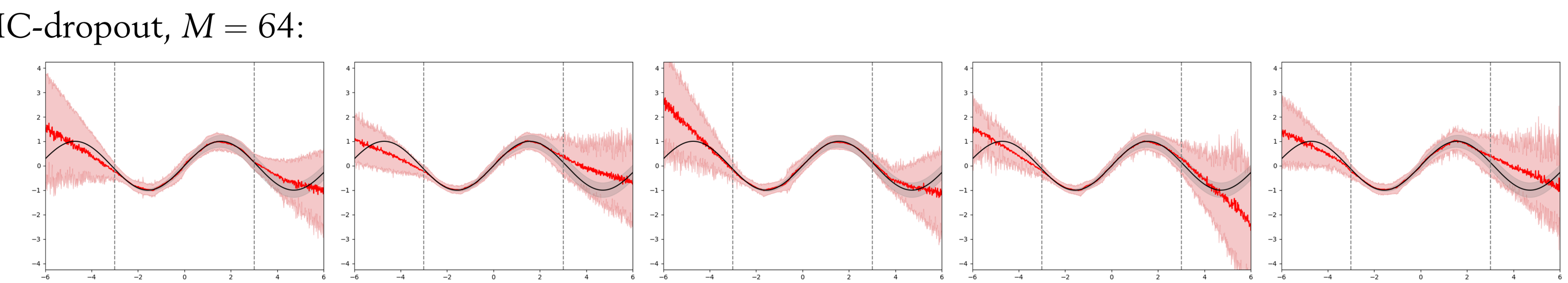
Illustrative toy problems - Regression



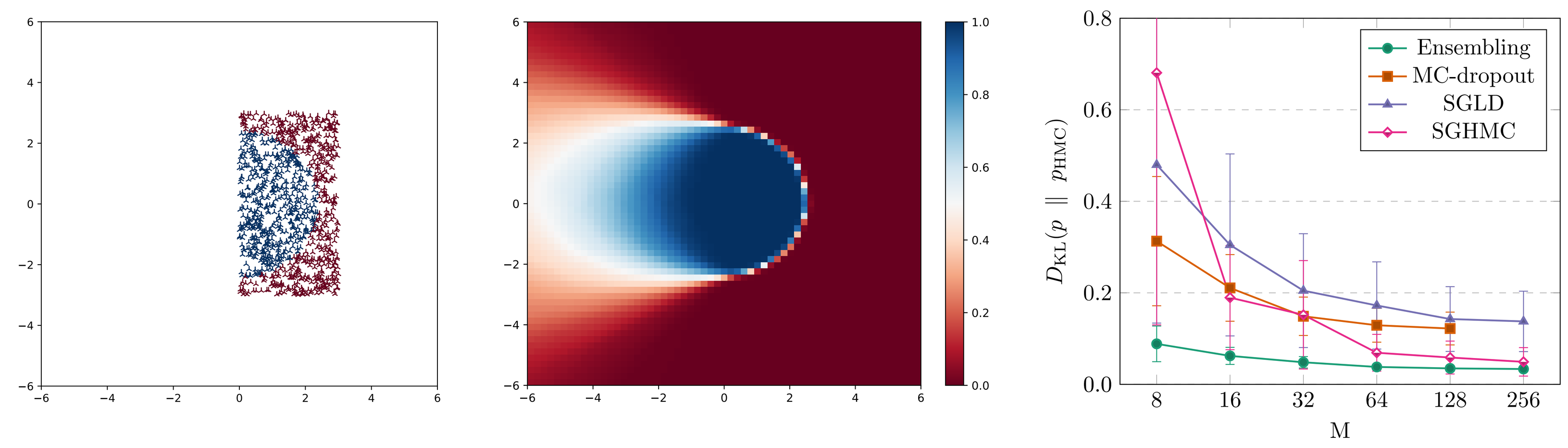
- Ensembling, $M = 64$:



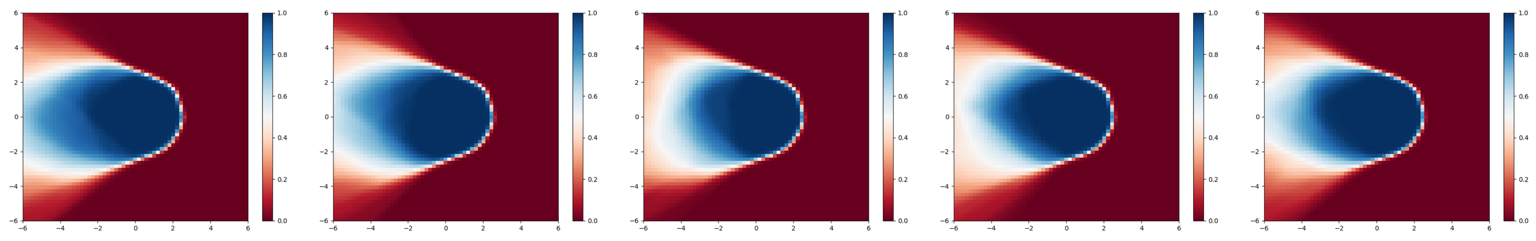
- MC-dropout, $M = 64$:



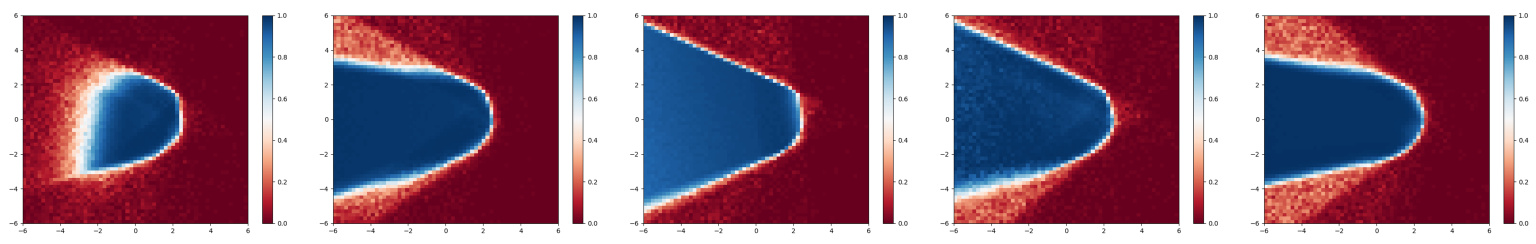
Illustrative toy problems - Classification



- Ensembling, $M = 64$:

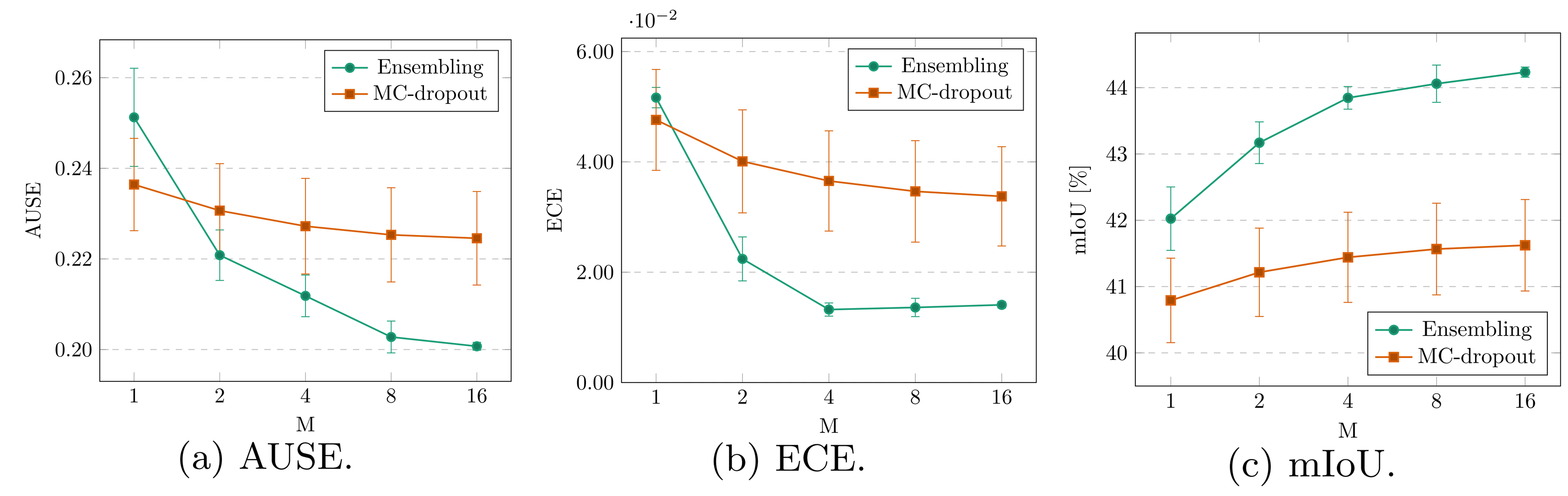


- MC-dropout, $M = 64$:



Street-scene semantic segmentation

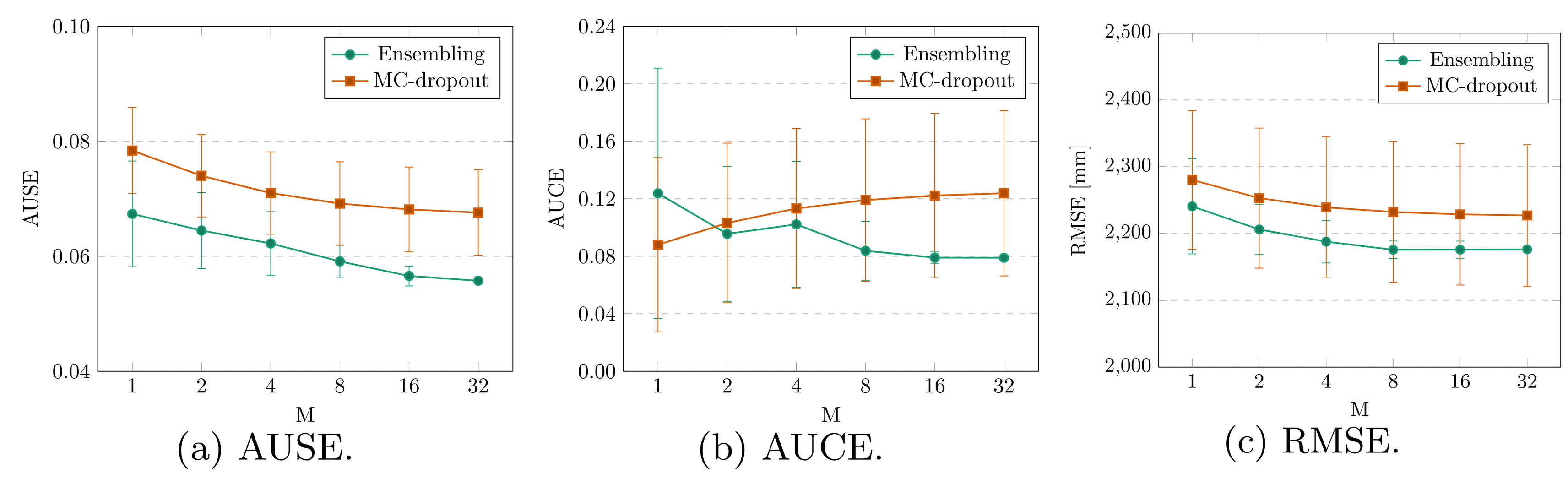
- Given an image $x \in \mathbb{R}^{h \times w \times 3}$, predict y of size $h \times w$, in which each pixel is assigned to one of C classes (road, car, etc.). Models are trained on **synthetic data** and evaluated on **real data**, testing robustness to **out-of-domain** inputs.
- Metrics for evaluation of uncertainty estimation quality:
  - **AUSE:** *relative* measure that reveals how well the estimated uncertainty can be used to sort predictions from worst (large true prediction error) to best.
  - **ECE:** *absolute* measure in terms of calibration. A well-calibrated model is not over-confident nor over-conservative.



(a) AUSE. (b) ECE. (c) mIoU.

Depth completion

- Given an image $x_{\text{img}} \in \mathbb{R}^{h \times w \times 3}$ and an associated *sparse* depth map, predict a *dense* depth map $y \in \mathbb{R}^{h \times w}$ of the scene. Models are trained on **synthetic data** and evaluated on **real data**, testing robustness to **out-of-domain** inputs.
- Metrics for evaluation of uncertainty estimation quality: **AUSE** and **AUCE** (generalization of ECE to the regression setting).



(a) AUSE. (b) AUCE. (c) RMSE.

Discussion & conclusion

- Required **training** scales linearly with M for ensembling, but this is not a major concern in most safety-critical applications, such as automotive. The main drawback of both methods is instead the computational cost at **test time** that scales linearly with M, affecting real-time applicability.
- Our work suggests that **ensembling** should be considered the new go-to method for *scalable* epistemic uncertainty estimation. We attribute its success to the ability to capture **multi-modality** in the posterior distribution $p(\theta|\mathcal{D})$.

fredrik.gustafsson@it.uu.se, martin.danelljan@vision.ee.ethz.ch, thomas.schon@it.uu.se