university of groningen / faculty of science and engineering / artificial intelligence

Not Just Seeing, Understanding What's Uncertain: Uncertainty in Deep Learning Stereo Depth Estimation

Bachelor's Project Thesis

Sven van Loon, s5218535, s.a.van.loon@student.rug.nl,
Supervisors: Dr M.A. Valdenegro Toro & MSc I.P. de Jong

Abstract: Stereo depth estimation is an important component for safety-critical systems such as autonomous vehicles, yet deep-learning–based approaches traditionally provide only point estimates without confidence measures. This thesis augments the Pyramid Stereo Matching Network (PSMNet) with a dual-output architecture to predict per-pixel disparity and variance, enabling the disentanglement of aleatoric (data-dependent) and epistemic (model-dependent) uncertainty. Three uncertainty quantification methods—Monte Carlo Dropout, Flipout, and Deep Ensembles—are implemented and trained on the DrivingStereo dataset at varying data fractions (10%, 50%, 100%). Models are evaluated based on their calibration, uncertainty disentanglement quality, and out-of-distribution detection. Deep Ensembles achieve the best in-distribution accuracy and calibration. Flipout tends to collapse its epistemic uncertainty, limiting its usefulness for OOD detection. Ensembles and MC Dropout detect OOD inputs via elevated epistemic uncertainty. Uncertainty estimates correlate strongly with prediction error for all methods.

1 Introduction

Stereo depth estimation has become an essential technique in the field of computer vision (Scharstein & Szeliski, 2002). Just as humans perceive depth through binocular vision by comparing the slightly different images seen by each eye, stereo vision systems use a similar principle to estimate depth from two cameras. By leveraging the disparity between images captured from slightly different viewpoints, stereo vision systems can reconstruct depth information essential for tasks such as autonomous navigation or robotics. Autonomous vehicles rely on stereo for obstacle localization (Geiger et al., 2012), surgical robots use stereo vision to guide precise movements during surgery (Probst et al., 2017), augmented-reality headsets trust disparity maps to superimpose virtual objects in 3-D space (Kanbara et al., 2000). In these settings, an incorrectly calculated depth can yield a collision, a laceration, or a misjudged maneuver that endangers road users. Knowing how sure a model is about its own prediction is therefore as important as the prediction itself.

Problem Statement In recent years, deep learning-based methods have revolutionized stereo matching by significantly improving the accuracy of depth estimation. Networks such as PSMNet (Chang & Chen, 2018) and ACVNet (Xu et al., 2022) leverage large datasets and complex architectures to learn features and matching strategies that outperform the traditional methods. Such networks excel at regressing per-pixel disparities but provide only point estimates, giving users no direct signal about confidence. Those models often exhibit a noticeable drop in performance when tested in unseen domains (Rao et al., 2023). Real-world scenes frequently contain variations in lighting or weather conditions that are unrepresented in the training data, showing the limitations of these data-driven approaches.

This presents a serious issue for safety-critical systems such as autonomous vehicles, where reliable depth estimation is essential for navigation and collision avoidance. For instance, an autonomous car must detect obstacles. It could lead to hazardous decisions if the stereo system produces un-

reliable depth predictions in foggy or low-light conditions without indicating the uncertainty. Embedding principled uncertainty allows systems to reason probabilistically: they can fuse multiple sensors, trigger fall-back planners, or ask for human intervention when uncertainty is high.

Such uncertainty can generally be categorized into two types: aleatoric uncertainty, which arises from the data, and epistemic uncertainty, which arises from the model (Hüllermeier & Waegeman, 2021). A more detailed explanation of these uncertainty types is provided in Section 2.

In this work, for each uncertainty quantification approach, we seek to address the following questions:

1. **Calibration**: How well do the predicted uncertainty intervals actually cover the true regression errors?

2. **Disentanglement**: How effectively can aleatoric and epistemic uncertainties be separated?

3. **Out-of-Distribution Detection**: How reliably do these methods identify inputs that deviate from the training distribution?

Contributions This thesis investigates the two types of uncertainties in the context of autonomous vehicles. Specifically, it takes the PSMNet architecture (Chang & Chen, 2018) and incorporates various uncertainty estimation techniques, including Monte Carlo Dropout (Gal & Ghahramani, 2016), Deep Ensembles (Lakshminarayanan et al., 2017), and Flipout (Wen et al., 2018). An overview of the example Ensemble framework can be found in Figure 1.1. These methods are evaluated based on how well they are calibrated, how the associated uncertainties behave across different scenarios and for their effectiveness in detecting out-of-distribution (OOD) samples. Additionally, each method is trained on varying amounts of data to assess how data size impacts uncertainty estimation.

Thesis Structure The structure of the thesis is as follows: Section 2 introduces relevant theory, Section 3 outlines the methodology used; Section 4 presents the experimental setup and results; and

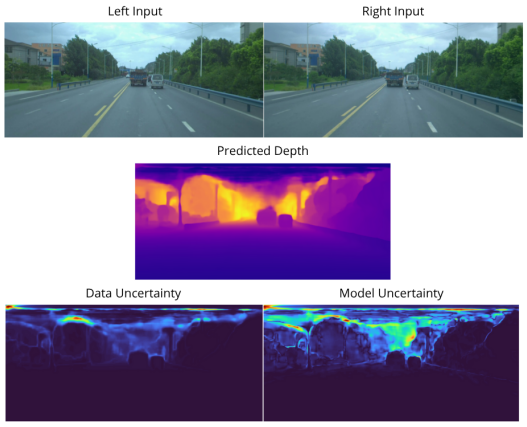


Figure 1.1: Thesis framework overview: stereo inputs are passed through a modified stereo depth network, the method produces depth map, data uncertainty map, and model uncertainty map.

Section 5 discusses the findings with suggestions for future work and Section 6 concludes the findings and thesis.

2 Background and Theory

Stereo vision infers the 3-D structure of a scene from two calibrated images taken from slightly different viewpoints (see Figure 2.1).

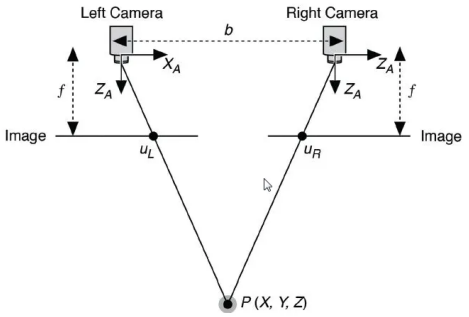


Figure 2.1: Pinhole-camera geometry for a rectified stereo pair. A 3-D point $P(X, Y, Z)$ projects to image coordinates (u_L, v) and (u_R, v) in the left and right views, separated by baseline b. Adapted from Eser, 2018.

Disparity Let (u, v) denote a pixel coordinate (column u, row v). A single 3-D point P will project to two different horizontal positions in a left-right stereo pair (see Figure 2.1). If those projections fall at pixel coordinates $u_L(u, v)$ in the left image and $u_R(u, v)$ in the right image, then their horizontal shift, called the **disparity**, can be calculated as

$$D(u, v) = u_L(u, v) - u_R(u, v), \tag{2.1}$$

where $u_L(u, v)$ and $u_R(u, v)$ are the horizontal pixel coordinates of the point's projections in the left and right images, respectively. By measuring disparity at each pixel, we obtain a dense map of stereo shifts that can be converted into depth estimates.

Disparity to Depth Conversion The modified PSMNet model predicts a per-pixel disparity map. Given camera intrinsics—focal length f and baseline B (distance between the two camera centers)—we convert disparity to depth via the well-known pinhole stereo formula (Hartley & Zisserman, 2003):

$$Z(u, v) \;=\; \frac{f \times B}{d(u, v)}. \tag{2.2}$$

Here, larger disparities correspond to closer objects, and vice versa.

2.1 Deep Stereo Networks and PSMNet

Stereo matching has evolved from handcrafted cost functions and global optimization into fully differentiable, end-to-end convolutional networks that learn to regress disparity directly from image pairs (Laga et al., 2020). In this subsection we first review the standard architecture of deep stereo methods and then describe the Pyramid Stereo Matching Network (PSMNet), the modified network used in this thesis.

Disparity Regression Modern CNN-based stereo models follow a four-stage pattern (Laga et al., 2020):

1. **Feature Extraction**
   Input images
   $$I_L, I_R \;\in\; \mathbb{R}^{h \times w \times 3},$$
   where I_L and I_R are the left and right RGB images, h and w denote the image height and width in pixels, and the third dimension 3 corresponds to the red, green, and blue color channels, are passed through a shared 2D backbone to yield feature maps
   $$F_L(u, v), F_R(u, v) \;\in\; \mathbb{R}^{c_f},$$
   where c_f is the number of feature channels per pixel in the output tensor.

2. **Cost-Volume Construction**
   For each disparity candidate
   $$d_c \in \{0, 1, \ldots, d_{\max} - 1\},$$
   where d_c is the horizontal shift in pixels and $d_{\max}$ is the maximum shift considered, we form
   $$C(u, v, d_c) = \mathrm{concat}\big(F_L(u, v),\, F_R(u - d_c,\, v)\big),$$
   yielding
   $$C \;\in\; \mathbb{R}^{h \times w \times d_{\max} \times 2\, c_f},$$
   where $C(u, v, d_c)$ is the concatenated feature vector at pixel (u, v) and shift d_c.

3. **Cost-Volume Regularization**
   A 3D CNN refines the cost volume
   $$C(u, v, d_c)$$
   over the axes (u, v, d_c) to produce a regularized cost
   $$C^{\mathrm{reg}}(u, v, d_c),$$
   where C^{reg} has the same shape as C. Architectures vary: stacked hourglass networks (Chang & Chen, 2018), 3-D UNets (Kendall et al., 2017), and recurrent GRUs (Du et al., 2021).

4. **Differentiable Disparity Read-Out**
   We convert the regularized costs into a probability mass function over d_c:
   $$p(d_c \mid u, v) = \sigma_s\big(-C^{\mathrm{reg}}(u, v, d_c)\big),$$
   where σ_s is the softmax operator over the disparity dimension, and then compute the soft-argmin (the expected disparity):
   $$\hat{D}(u, v) = \sum_{d_c=0}^{d_{\max}-1} d_c \; p(d_c \mid u, v),$$
   yielding a continuous disparity map $\hat{D}$.

2.1.1 Pyramid Stereo Matching Network (PSMNet)

We adopt PSMNet (Chang & Chen, 2018) as the core stereo matching backbone for this thesis due to its effective balance between accuracy and computational efficiency.

Architecture Overview The architecture comprises a ResNet-based encoder with dilated convolutions, spatial pyramid pooling for multi-scale context, and a 4D cost volume built from fused features. This cost volume is refined using a 3D CNN with stacked hourglass modules, and final disparity maps are predicted via soft-argmin regression for end-to-end training.

Selection Reason PSMNet was chosen for the following reasons:

1. **State-of-the-Art Performance:** PSMNet achieves competitive accuracy on benchmark datasets.
2. **Community Adoption:** PSMNet has become a widely used baseline in stereo research.
3. **Open-Source Availability:** The PSMNet implementation is publicly available under an open-source license on GitHub.

2.2 Uncertainty in Deep Learning

In the context of deep learning, uncertainty refers to the model's lack of confidence or reliability in its predictions. In literature (Hüllermeier & Waegeman, 2021), such uncertainty can generally be categorized into two types:

Aleatoric Uncertainty (AU), which is inherent to the data and arises from noisy inputs, such as measurements from LIDAR, and *cannot be reduced* even with more data.

Epistemic Uncertainty (EU), which comes from the model itself and arises from a lack of knowledge due to limited training data or model capacity, and *can be reduced* by collecting more data or improving the model.

Combined, these two components are often referred to as the model's total uncertainty (TU).

In stereo depth estimation, Wang et al. (2022) introduce an end-to-end stereo matching framework that predicts per-disparity Normal Inverse-Gamma distribution parameters. Mehltretter (2022) propose a Bayesian neural network trained via variational inference with a probabilistic loss to jointly predict disparity and both aleatoric and epistemic uncertainties. However, research on uncertainty estimation in stereo depth estimation remains limited.

Disentanglement Distinguishing between these two types of uncertainty is important because it helps in identifying whether the unreliability stems from the data or from the model. However, recent work has raised concerns about whether such disentanglement is truly achieved in practice. Wimmer et al. (2023) proves that conditional entropy and information-theoretic approaches to uncertainty quantification methods are fundamentally incoherent. Building on this insight, de Jong et al. (2024) provide experimental evidence that information-theoretic and the Gaussian Logits approaches fail to reliably separate aleatoric from epistemic uncertainty in classification settings. Their findings suggest that EU and AU are contaminated by one another. However, these critiques concern classification, where both the methodology and the distributions involved differ. In regression contexts—where one typically employs a variance-based decomposition (via the law of total variance described below) rather than entropy-based measures—the practical implications of these incoherence results remain less clear, and regression models may avoid some of the pitfalls identified in the classification. A recent theoretical analysis by Bülte et al. (2025) introduces an axiomatic framework specifically for regression, demonstrating that variance-based decompositions always remain nonnegative but may ignore epistemic uncertainty in the predicted variance itself. These insights directly motivate our investigation of how to interpret uncertainty in stereo depth estimation.

Disentanglement in Regression In the regression setting, we assume a model that, for each input vector x (e.g. the features describing a data point) and for each stochastic forward-pass or ensemble member $i \in \{1, \ldots, T\}$, produces both a predictive mean $\mu_i(x)$ and predictive variance $\sigma_i^2(x)$. Here, y denotes the scalar output (target) we wish to pre-

dict. These T samples from the (approximate) posterior are then combined into a single Gaussian approximation of the predictive distribution of y given x:

$$p(y \mid x) \approx \mathcal{N}\big(\mu_*(x),\, \sigma_*^2(x)\big), \tag{2.3}$$

where Valdenegro-Toro & Mori (2022) show that the predictive mean and variance are given by

$$\mu_*(x) = \frac{1}{M}\sum_{i=1}^{M} \mu_i(x), \tag{2.4}$$

$$\sigma_*^2(x) = \frac{1}{M}\sum_{i=1}^{M} \big[\sigma_i^2(x) + \mu_i(x)^2\big] - \mu_*(x)^2. \tag{2.5}$$

Applying the law of total variance, Valdenegro-Toro & Mori (2022) decompose $\sigma_*^2(x)$ into aleatoric and epistemic components:

$$\sigma_*^2(x) = \underbrace{\mathbb{E}_i\big[\sigma_i^2(x)\big]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_i\big[\mu_i(x)\big]}_{\text{epistemic}}. \tag{2.6}$$

Loss Function To train the variance head $\sigma^2(x)$ one typically minimizes the Gaussian negative log-likelihood (Valdenegro-Toro & Mori, 2022):

$$\mathcal{L}_{\mathrm{NLL}}\big(y_{\mathrm{gt}},\, \mu(x),\, \sigma^2(x)\big) = \frac{1}{2}\log \sigma^2(x) + \frac{\big(\mu(x) - y_{\mathrm{gt}}\big)^2}{2\,\sigma^2(x)} \tag{2.7}$$

where y_{gt} indicates the ground-truth value.

This loss encourages the model to predict lower variance when its mean predictions are close to the ground truth, and higher variance when the errors are larger.

2.3 Uncertainty Methods

To capture both uncertainties, we compare three approaches based on Valdenegro-Toro & Mori (2022):

1. **MC Dropout:** Monte Carlo Dropout approximates Bayesian inference by keeping dropout at test time.

2. **Flipout:** Flipout layers employ variational inference to learn a distribution over weights instead of point estimates. During both training and inference, pseudo-independent weight perturbations are generated by applying random sign flips to sampled weights.

3. **Deep Ensembles:** Ensembles approximate Bayesian model averaging by training T independent models with identical architectures but different weight initializations.

Inference At test time, T stochastic forward passes or T forward passes through each ensemble member produce

$$\{\hat{\mu}_t,\, \hat{\sigma}_t^2\}_{t=1}^{T}, \tag{2.8}$$

whose sample mean and variance approximate the predictive distribution.

Other Methods Concrete Dropout (Gal et al., 2017) extends standard dropout by learning input-dependent dropout probabilities. Valdenegro-Toro & Mori (2022) further propose a DropConnect variant, which randomly zeroes individual weights at each pass to induce parameter uncertainty. Due to the training computational and memory costs, these methods were not employed in our experiments.

2.4 Evaluation of Uncertainty Quantification

A good depth-estimation model must not only minimise error but also communicate when its answers are likely to be wrong. We therefore evaluate three complementary properties—calibration and error–uncertainty interaction—all computed at the per-pixel level on test images.

Calibration Calibration evaluates how faithfully a model's uncertainty estimates correspond to its actual prediction errors. In a perfectly calibrated regression model, if we construct an $\alpha = 90\%$ predictive interval around each prediction, then approximately 90% of the true targets should fall within those intervals.

Good calibration in stereo-depth estimation ensures that reported confidence levels correspond to actual depth-error frequencies and that uncertainty rankings align with pixel-wise error. This is critical for downstream applications (e.g. obstacle avoidance), where overconfident or miscalibrated depth predictions can lead to unsafe decisions.

One way to investigate calibration is to construct a reliability plot:

1. **Confidence** Define an evenly–spaced set of target confidences $\mathcal{S}_\alpha = \{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$.

2. **Per–pixel prediction intervals.** For every pixel i with predictive mean μ_i and EU standard deviation σ_{epi}, construct the two–sided α-level interval

$$I_i(\alpha) = \left[\mu_i - z_{\alpha/2}\,\sigma_{\text{epi}},\ \mu_i + z_{\alpha/2}\,\sigma_{\text{epi}}\right], \quad (2.9)$$

   where $z_{\alpha/2}$ is the standard-normal quantile.

3. **Coverage** The observed coverage at confidence α is

$$\text{cov}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{y_i^{\text{gt}} \in I_i(\alpha)\right\}, \quad (2.10)$$

   with y_i^{gt} the ground-truth value.

4. **Reliability Plot** Plot the pairs $\big(\alpha, \text{cov}(\alpha)\big)$. Perfect calibration lies on the diagonal; curves below it indicate over-confidence.

Calibration of predicted uncertainties can be summarized by a scalar score, one common metric is the Calibration Error (CE):

$$\text{CE} = \frac{1}{|\mathcal{S}_\alpha|}\sum_{\alpha\in\mathcal{S}_\alpha} \big|\text{cov}(\alpha) - \alpha\big|, \quad (2.11)$$

Another common metric is the Maximum Calibration Error (MCE):

$$\text{MCE} = \max_{\alpha\in\mathcal{S}_\alpha} \big|\text{cov}(\alpha) - \alpha\big|. \quad (2.12)$$

MCE highlights the largest miscalibration, making it sensitive to extreme calibration errors in any single confidence region (Guo et al., 2017).

Sparsification Curves In dense-prediction tasks such as stereo depth estimation, sparsification curves provide a complementary perspective: by ranking pixels by predicted uncertainty and progressively removing the most uncertain ones, one observes whether the remaining error decays rapidly, indicating that high-uncertainty pixels indeed coincide with high-error regions (Poggi et al., 2020). They are constructed using the following algorithm:

1. Rank all pixels by predicted uncertainty (largest uncertainties first).

2. Progressively remove the top $p\%$ most-uncertain pixels and recompute a summary error metric on the remainder; sweep p from 0 to 100%.

3. Plot error versus retained-pixel fraction.

Error–Uncertainty Correlation To assess how well predicted uncertainties align with actual prediction errors, one can compute the correlation between the estimated uncertainty σ_i and the absolute error $e_i = |y_i^{\text{gt}} - \hat{\mu}_i|$ across all pixels. Common metrics include the Pearson correlation coefficient:

$$\rho = \frac{\sum_i (\sigma_i - \bar{\sigma})(e_i - \bar{e})}{\sqrt{\sum_i (\sigma_i - \bar{\sigma})^2}\,\sqrt{\sum_i (e_i - \bar{e})^2}}. \quad (2.13)$$

2.5 OOD Detection

Out-of-Distribution (OOD) refers to inputs at test time that lie outside of the training data distribution. In other words, if the model was trained on images of sunny scenes, a rainy scene would be OOD. Detecting OOD inputs is crucial because models tend to make overconfident predictions on data they've never seen.

Good OOD detectors help avoid catastrophic failures by allowing models to say "I don't know" when faced with unfamiliar data, handing control back to downstream safety mechanisms or humans.

Epistemic Uncertainty for OOD Detection Epistemic uncertainty measures the model's lack of knowledge and is high in regions of the input space that were sparsely or never seen during training (Kendall & Gal, 2017). This makes it a natural indicator for out-of-distribution (OOD) inputs: since the model has not learned reliable representations for such data, the variance across stochastic or ensemble predictions—and hence the epistemic term—tends to increase. Empirically, deep ensembles have been shown to assign significantly higher predictive uncertainty to OOD examples compared to in-distribution samples (Lakshminarayanan et al., 2017). In contrast, Valdenegro-Toro & Mori

(2022) showed on a toy dataset that Flipout's epistemic uncertainty estimates are unreliable for detecting OOD inputs.

3 Methods

In this section, we describe our modified PSMNet with dual heads for mean disparity and variance prediction, present three uncertainty methods (MC-Dropout, Flipout, and ensembles), outline dataset splits for DrivingStereo (ID) and vKITTI 2 (OOD), and define our metrics for depth accuracy, calibration, uncertainty quality, and OOD detection.

3.1 Modified Architecture

Our base stereo matching backbone is PSMNet (Chang & Chen, 2018), which processes a rectified stereo pair through two weight-sharing CNN streams, a Spatial Pyramid Pooling (SPP) module, and a 3D cost-volume regularizer. To disentangle aleatoric from epistemic uncertainty, we augment PSMNet with a second output "head" that predicts per-pixel variance. We extend the original architecture with two parallel output heads:

- A *disparity head* (standard PSMNet regressor), and
- A *variance head* (parallel regressor with softplus activation) that estimates AU.

The architecture can be seen in Figure 3.1.

Negative Gaussian Log-Likelihood After removing all the constants from Equation 2.7, we obtain the following loss function, which we minimize to jointly supervise both the mean and variance heads:

$$\mathcal{L}_{NLL}\big(d_i^{\text{gt}}, \hat{d}_i, \hat{\sigma}_i^2\big) = \log\big(\hat{\sigma}_i^2 + \epsilon\big) \; + \; \frac{\big(\hat{d}_i - d_i^{\text{gt}}\big)^2}{\hat{\sigma}_i^2 + \epsilon}, \tag{3.1}$$

where d_i^{gt} is the ground-truth disparity for pixel i, $\hat{d}_i$ is the predicted mean, $\hat{\sigma}_i^2$ is the predicted variance, and ϵ is a small constant for numerical stability.

This loss encourages the model to predict lower variance when its mean predictions $\hat{d}_i$ are close to the ground truth d_i^{gt}, and higher variance $\hat{\sigma}_i^2$ when the errors are larger.

3.2 Uncertainty Methods

To capture both uncertainties, we compare three approaches:

MC-Dropout We add a dropout layer with a dropout rate of 0.3 immediately before the final learnable layer, and keep it active at test time. During inference, we perform $T = 20$ stochastic forward passes through the network, each time sampling a different dropout mask. This yields a set of T predictions $\{\hat{d}_t, \hat{\sigma}_t^2\}_{t=1}^T$, from which we can calculate both epistemic and aleatoric uncertainty.

Flipout The final learnable layer is implemented as a Flipout layer and trained using variational inference. During inference, we perform $T = 20$ stochastic forward passes, each time sampling a new set of weight perturbations. This produces T predictions $\{\hat{d}_t, \hat{\sigma}_t^2\}_{t=1}^T$, from which we can calculate both epistemic and aleatoric uncertainty.

Ensembles We train an ensemble of $M = 3$ independent models, each sharing the same modified PSMNet dual-head network architecture without any added special layers or modifications, but with different random weights initializations. At test time, we perform one deterministic forward pass through each ensemble member, yielding predictions $\{\hat{d}_m, \hat{\sigma}_m^2\}_{m=1}^M$, from which we can calculate both epistemic and aleatoric uncertainty.

3.3 Hyperparameters

All training and architectural hyperparameters used in the experiments (learning rates, batch sizes, number of epochs, etc.) are listed in Appendix D.

3.4 Datasets

All models were trained and validated on DrivingStereo (Yang et al., 2019) dataset and evaluated on both DrivingStereo as In-distrubtion (ID) and Virtual Kitti 2 (Cabon et al., 2020) as Out-of-distribution (OOD)

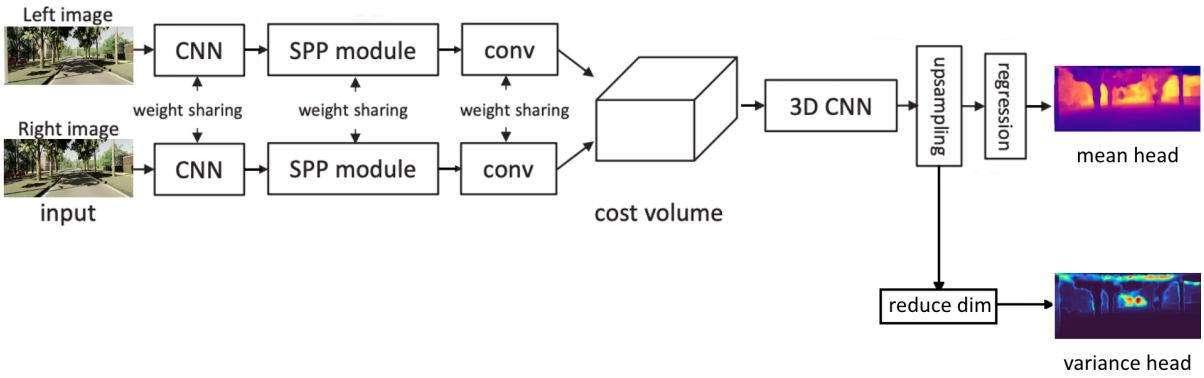


Figure 3.1: Architecture of the augmented PSMNet with 2 output heads.

DrivingStereo DrivingStereo (Yang et al., 2019) is a large-scale real-world sparse stereo driving dataset comprising 181 638 (of which 7 751 are test set) rectified stereo pairs captured across diverse urban and highway scenes.

Validation set Since the DrivingStereo dataset does not provide a predefined validation set, we constructed one by randomly sampling 9,158 stereo image pairs from the original training set. These pairs were held out and excluded from training to detect overfitting. After this split, 165279 stereo pairs remained for training. This same validation set was used consistently across all training data sizes (10%, 50%, and 100%).

Sub-sampling for 10%, 50% and 100% The full train pool (100 %) is the entire 165279-pair training set. To evaluate uncertainty and accuracy under varying amounts of supervision, we created 50% and 10% training subsets by randomly sampling from the 100% set. This sub-sampling was done independently for each subset, and the same validation set was used across all configurations.

Virtual KITTI 2 (vKITTI 2) Virtual KITTI 2 (Cabon et al., 2020) is a photorealistic synthetic stereo dataset simulated under varied lighting and weather conditions. We used 2 000 stereo pairs from this dataset as an out-of-distribution (OOD) test set to assess each method's uncertainty estimation on unseen domains.

Virtual KITTI 2 introduces a domain shift compared to DrivingStereo due to its synthetic nature and perfectly accurate ground-truth depth. Unlike the real-world images in DrivingStereo, vKITTI 2 contains no sensor noise, inconsistent lighting, or occlusions typical of real driving scenes. We expect models to predict lower aleatoric uncertainty on vKITTI 2, reflecting the absence of real-world ambiguity and noise present in the DrivingStereo dataset.

Split	# Stereo Pairs
Train (10% subset)	16,528
Train (50% subset)	82,640
Train (100% subset)	165,280
Validation (DrivingStereo)	9,158
ID Test (DrivingStereo)	1,000
OOD Test (vKITTI 2)	1,000

Table 3.1: Dataset splits and sizes used in our experiments.

Dataset Split As summarized in Table 3.1, we train on three DrivingStereo subsets (10%, 50%, and 100%), validate on 9,158 pairs, and evaluate on 1,000 in-distribution and 1,000 out-of-distribution (vKITTI 2) stereo pairs.

3.5 Evaluation Metrics

This section formalises how we score depth accuracy, uncertainty quality, and the relationship between the two.

Depth Accuracy To evaluate the accuracy of predicted depth maps, we use two standard regression metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics are computed between the predicted and ground-truth disparity maps and are reported over all valid pixels within the maximum disparity threshold of 224 pixels.

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|d_i^{\text{gt}} - \hat{d}_i\right| \tag{3.2}$$

  where d_i^{gt} is the ground-truth disparity, $\hat{d}_i$ is the predicted disparity, and N is the number of valid pixels.

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i^{\text{gt}} - \hat{d}_i)^2} \tag{3.3}$$

  RMSE penalizes larger errors more than MAE and is thus more sensitive to outliers.

In addition to these quantitative metrics, we also conduct a qualitative analysis of the predicted depth maps. This includes visual inspection that may not be fully reflected in the scalar metrics.

Uncertainty Quality To assess the quality of uncertainty estimates produced by each model, we employ a combination of quantitative metrics, statistical correlation analysis, and visual tools. These evaluations help determine how well the predicted uncertainties align with actual prediction errors and whether the uncertainty estimates are reliable and informative. Specifically, we consider:

1. **Calibration Error:** Measures the discrepancy between predicted uncertainty and actual prediction error. A well-calibrated model will produce uncertainty estimates that match the observed error distribution. We compute the Calibration Error described in Section 2.

2. **Reliability Plot:** A visual diagnostic that plots the actual error against predicted uncertainty across confidence values. Perfect calibration corresponds to a diagonal line; deviations from this line indicate under- or over-confidence.

3. **Pearson Correlation Coefficient:** We compute the Pearson correlation between predicted uncertainty and absolute error to quantify the strength and direction of their linear relationship. A strong positive correlation indicates that higher uncertainty reliably corresponds to higher prediction error, which is desirable for uncertainty estimation.

4. **Sparsification Plots:** These plots evaluate how well uncertainty estimates can be used to identify and remove high-error pixels. By gradually removing pixels with the highest predicted uncertainty, we plot the resulting MAE or RMSE against the retained fraction of pixels. A sharper drop in error with fewer pixels retained indicates more effective uncertainty estimation.

5. **Disentanglement:** To study how epistemic and aleatoric uncertainties behave under varying sizes of training set, we compare the distribution of predicted uncertainties across different training data sizes (10%, 50%, and 100%). This analysis helps see how well the models disentangle AU and EU.

6. **Qualitative Analysis:** We include visualizations of predicted uncertainty maps alongside the corresponding disparity predictions. These visual comparisons help assess whether uncertainty is meaningfully concentrated in ambiguous regions such as object boundaries, occlusions, or low-texture areas.

OOD Detection We evaluate OOD detection via two complementary protocols:

1. **Unseen synthetic data (vKITTI 2).** We apply our trained models (trained and validated only on DrivingStereo) to the Virtual KITTI 2 dataset (Test B). Since vKITTI 2 scenes were never seen during training, we expect an increase in EU relative to in-distribution (DrivingStereo) inputs.

2. **Left–right image swap.** As an in-distribution perturbation, we take DrivingStereo stereo pairs and swap the left and right images before feeding them to the network. Although these are still DrivingStereo

images, the incorrect correspondence should confuse the model and produce elevated epistemic uncertainty. We compare the distribution of EU on swapped versus correctly ordered pairs to quantify sensitivity to this form of internal mismatch.

In both settings, a reliable OOD detector should produce a clear separation between the EU distributions of nominal versus OOD (or perturbed) inputs. The epistemic and aleatoric uncertainty distributions are compared.

4 Experiments and Results

In this section we evaluate the uncertainty quantification methods on two datasets (DrivingStereo and Virtual KITTI2), both in-distribution (ID) and out-of-distribution (OOD), and analyze their calibration, OOD detection capability, and the relationship between predicted uncertainties and actual errors. All uncertainties are reported as standard deviations.

4.1 Qualitative Analysis

In the predicted depth panel of Figure 4.1, we observe that the Ensemble tends to hallucinate content in the upper region of the image. We believe this happens because DrivingStereo provides very few valid depth measurements in that region leading to uninformative gradients and unintended weight updates. Because the DrivingStereo dataset is so sparse, the resulting error map is itself quite sparse, and we observe that those pixels with valid depth values that also exhibit higher error tend to show correspondingly higher uncertainty—especially epistemic uncertainty. In fact, epistemic and aleatoric uncertainties are often colocalized: AU remains concentrated along object boundaries (including the boundaries of objects imagined/hallucinated by the model), whereas EU is especially pronounced inside those hallucinated objects and in regions where the model predicts surfaces or objects that are not actually present.

Figure 4.2 presents results from the Deep Ensemble model tested on an out-of-distribution (OOD) dataset captured under rainy conditions. The error map reveals that the model's depth estimates become increasingly inaccurate for more distant objects—an effect compounded by the fact that the model was trained on data with a maximum range of 80 meters, so it never predicts beyond that limit. This also explains the difference between ground-truth depth and predicted depth maps. Under these rainy conditions, synthetic raindrops on the lens are sometimes interpreted as scene geometry (most notably in the middle lower region of the predicted depth), producing spurious "objects" that the model hallucinates. These artifacts are predominantly captured by spikes in epistemic uncertainty (EU), reflecting the model's lack of confidence inside and around these falsely detected regions, while aleatoric uncertainty (AU) remains focused on the true and hallucinated object boundaries (including the edges of the raindrop-induced artifacts) and generally increases with depth. In these areas of joint elevation, AU delineates the object outlines and EU highlights the interior of hallucinated structures, together providing a clear signature of where and why the model fails under challenging, unseen conditions.

Visualizations for other methods and their comparison can be seen in Appendix A.

4.2 Depth Prediction

Table 4.1 summarizes the mean absolute error (MAE), root-mean-square error (RMSE) for each method at full train set sizes (100%). On the ID DrivingStereo set, all three approaches achieve sub-half-metre MAE at full resolution, with the Ensemble slightly outperforming MC Dropout and Flipout (MAE = 0.469 m vs. 0.494 m and 0.484 m, respectively). Under OOD conditions (Virtual KITTI2), errors increase by around 8m, but Ensemble achieves the lowest MAE (8.305 m) and RMSE (16.350 m) among the methods. Masking the near-field (eg. only considering the depth the model was trained to output) reduces MAE by roughly a factor of two, with the Ensemble still maintaining the best performance.

4.3 Calibration

We assess how well the epistemic uncertainty bounds match empirical errors via the reliability plot in Figure 4.3. Ideally, a model with perfect calibration would achieve coverage equal to the

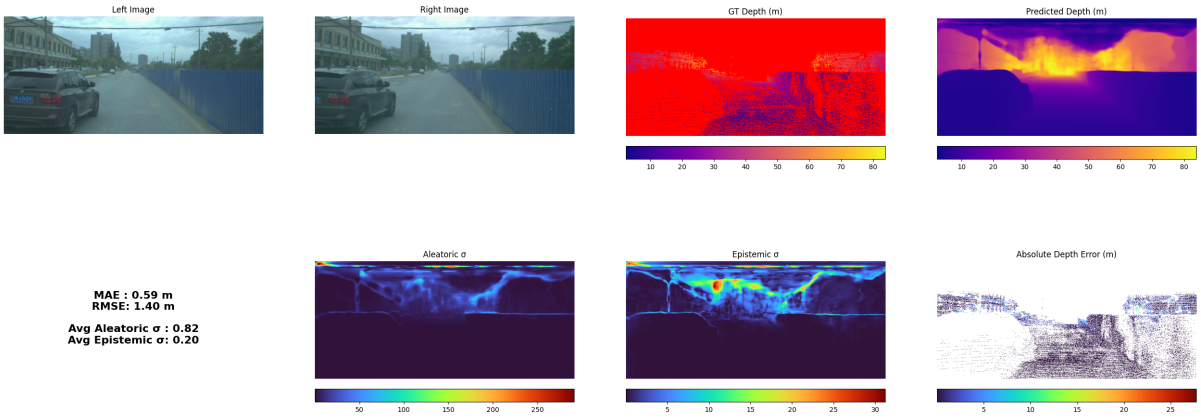


Figure 4.1: Ensemble 100%: Qualitative results (DrivingStereo) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 0.59 m, RMSE = 1.40 m, average AU = 0.82 and average EU = 0.20.

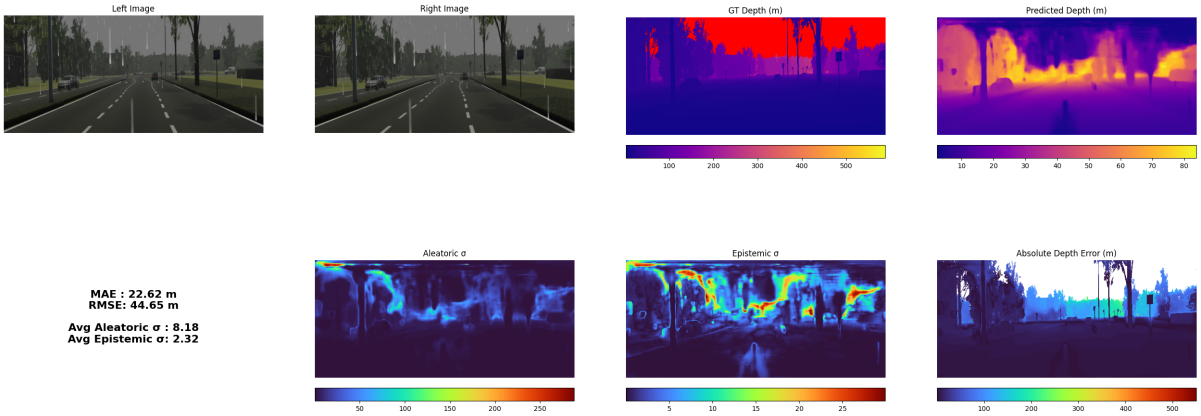


Figure 4.2: Ensemble 100%: Qualitative results (Virtual Kitti 2) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 22.62 m, RMSE = 44.65 m, average AU = 8.18 and average EU = 2.32.

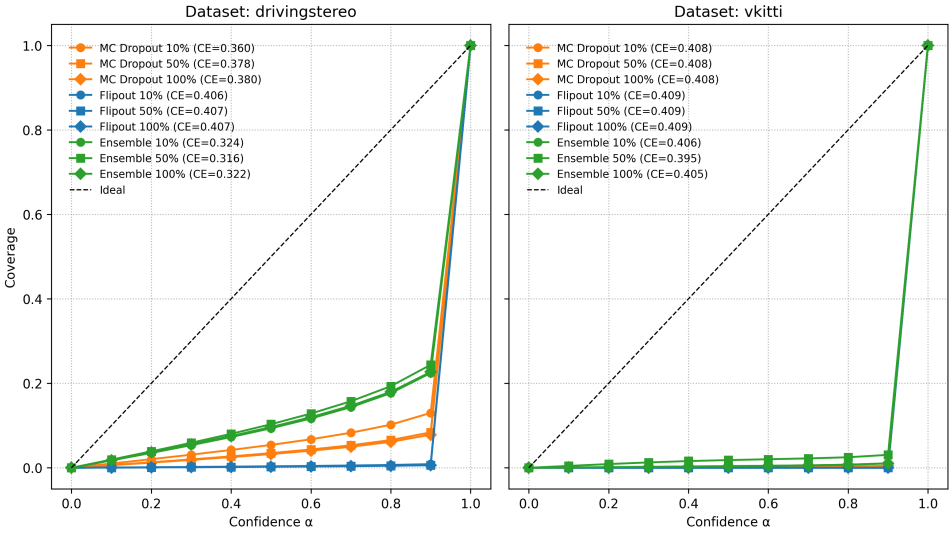


Figure 4.3: Reliability diagrams for DrivingStereo (left) and Virtual Kitti 2 (right), showing coverage versus confidence α for MC Dropout (orange), Flipout (blue), and deep ensemble (green) models trained on 10%, 50% and 100% of the training set (legend reports calibration error CE for each method), the dashed diagonal indicates perfect calibration.

Method	Error [m]	
	MAE	RMSE
DrivingStereo		
MC Dropout 100%	0.494	1.102
Flipout 100%	0.484	1.087
Ensemble 100%	**0.469**	**1.043**
VirtualKITTI2 – OOD		
MC Dropout 100%	8.406	16.458
Flipout 100%	8.821	17.765
Ensemble 100%	**8.305**	**16.350**
VirtualKITTI2 – OOD (masked)		
MC Dropout 100%	5.255	7.242
Flipout 100%	5.406	7.800
Ensemble 100%	**5.164**	**7.086**

Table 4.1: Mean absolute error (MAE) and root-mean-square error (RMSE) for the DrivingStereo (in-distribution) and VirtualKITTI2 (out-of-distribution) datasets, including masked on VirtualKITTI2. Bold values denote the lowest error achieved in each setting.

nominal confidence (Coverage(α) = α). In DrivingStereo (left panel), ensembles yield the lowest calibration error ($CE \approx 0.32$), followed by MC Dropout ($CE \approx 0.38$) and Flipout ($CE \approx 0.41$). All methods are overconfident. On Virtual KITTI2 (right panel), ensembles remain best calibrated ($CE \approx 0.40$), while MC Dropout and Flipout exhibit almost identical but very poor calibration ($CE \approx 0.41$).

4.4 OOD detection

An effective OOD detector should assign higher epistemic uncertainty to OOD (Virtual KITTI 2 or switched left and right) samples than to in-distribution DrivingStereo frames.

DrivingStereo vs. Virtual KITTI 2 Figure 4.4 shows the distributions of AU and EU for all methods. In DrivingStereo, aleatoric uncertainties dominate and are broadly similar across methods, while epistemic uncertainties are much smaller, with Flipout predicting near zero EU. Conversely, on Virtual KITTI2 both types of uncertainty increase, especially epistemic uncertainty under Flipout and Ensembles indicating successful

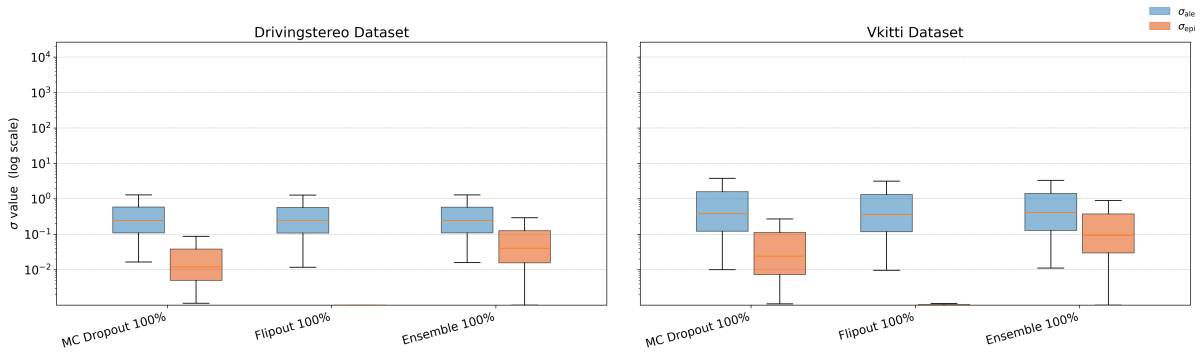


Figure 4.4: Box plots (log-scale on the y–axis) demonstrating distributions of per-pixel AU (blue) and EU (orange) uncertainty for MC Dropout, Flipout, and Deep Ensemble models trained on 100%. The DrivingStereo distributions are on the left and Virtual KITTI 2 are on the right.

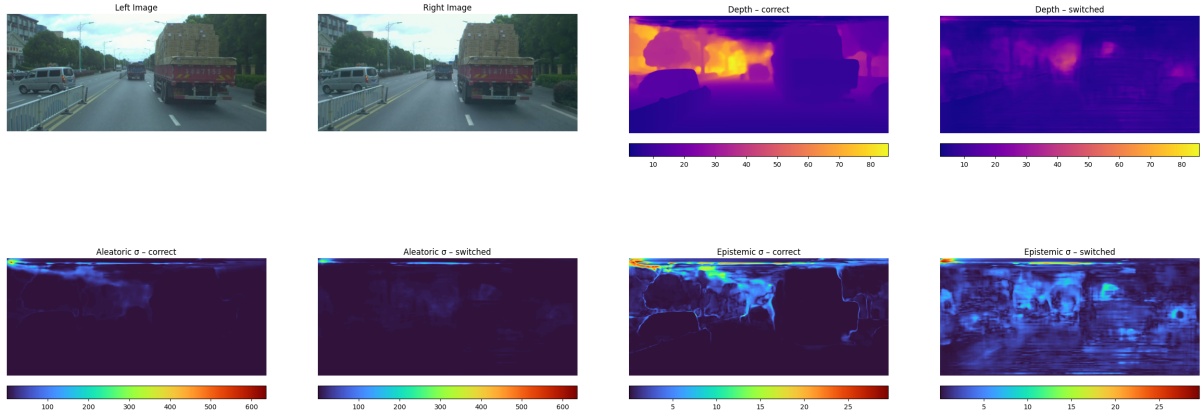


Figure 4.5: Ensemble 100%: Qualitative Results (DrivingStereo) from left-right input swap test. Panels labeled "correct" show properly paired stereo inputs, while panels labeled "switched" were obtained by swapping the left and right inputs. As desired, epistemic uncertainty (EU) increases in the switched scenario.

OOD detection.

Left–right image swap The swapped inputs produce wildly incorrect depth maps (see Fig. 4.5). It can be clearly seen that the EU rises drastically throughout the whole image, indicating good OOD detection.

Results of the left–right image swap for the remaining methods are provided in Appendix B.

4.5 10% vs. 50% vs. 100%

In this section we examine how the fraction of training data—10%, 50% or the full 100% of DrivingStereo affects the resulting uncertainty distributions. We expect the EU to decrease and AU to stay the same as we increase the training set size.

Aleatoric Uncertainty. As seen in Figure 4.6, across all three methods, the distribution of AU for DrivingStereo remains essentially unchanged as the training-set fraction grows from 10% to 100%. This stability reflects that noise is learned even from a small subset of the data.

Epistemic Uncertainty (In-Distribution). On DrivingStereo (left half of each subplot), EU consistently decreases with more training data for MC Dropout and Flipout. For MC Dropout

(Fig.4.6a), the median EU falls slightly from 10% to 100%, indicating that additional data helps the model resolve its own uncertainty on familiar scenes. Deep Ensembles (Fig.4.6b) show very little drop. Flipout's EU is near zero regardless of data fraction (Fig. 4.6c), still showing little drop.

4.6 Are higher uncertainties associated with higher error?

Supplementary bar-chart panels for this analysis are provided in Appendix C.

Sparsification Plot Sparsification curves in Figures 4.7a and 4.7b quantify how MAE decreases as pixels with highest uncertainty are progressively removed. In both datasets, ranking by each type of uncertainty leads to error reduction. However, on the Virtual KITTI 2 dataset, MC Dropout exhibits a handful of low-ranked pixels that incur larger errors. Because the AU is much larger than the EU, meaning aleatoric uncertainty "dominates" the total uncertainty, the total (green) and aleatoric curves coincide almost exactly, rendering the TU line invisible in the plot.

Correlation Analysis When comparing the error–uncertainty correlations to the error–depth correlations in Table 4.2, we find that uncertainty is an even stronger indicator of per-pixel error than the predicted depth itself on ID dataset. On DrivingStereo, Flipout's aleatoric and total uncertainty both correlate with error at $r = 0.635$, achieving the strongest correlation between uncertainty and error. Its correlation between error and predicted depth is noticeably lower ($r = 0.578$).

Under domain shift on VirtualKITTI2 (second part of Table 4.2), we observe that the best indicator for error is the predicted depth values across all three methods. Among uncertainty estimates, the highest correlation is achieved by MC Dropout's epistemic uncertainty ($r = 0.342$).

4.7 Uncertainty against depth

Across all methods in the DrivingStereo, aleatoric uncertainty exhibits a positive correlation with true depth (see Tab. 4.2), reflecting the fact that farther pixels inherently produce noisier disparity estimates. Epistemic uncertainty, by contrast, shows weaker correlation with depth (e.g. $r = 0.257$ for Flipout 100% on DrivingStereo), suggesting that model uncertainty is not simply a function of distance but rather of appearance unfamiliarity.

Supplementary bar-chart panels for this analysis are provided in Appendix C.

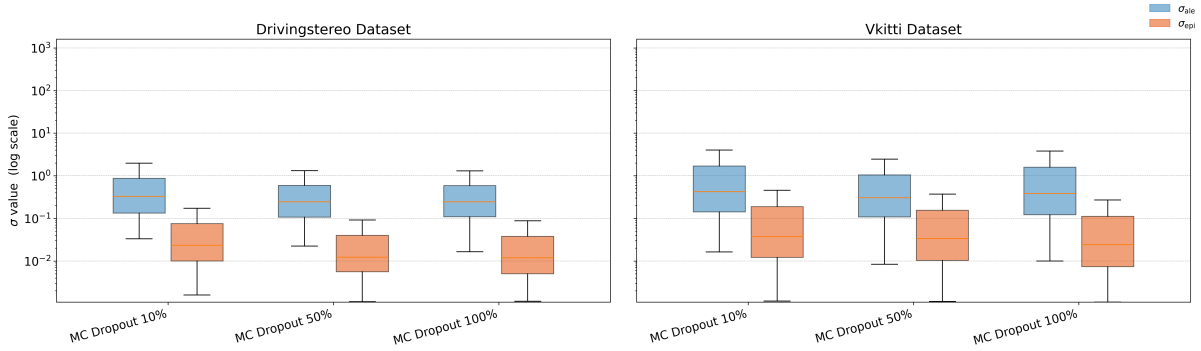
5 Discussion

In this work, we have evaluated three uncertainty extensions of PSMNet—MC Dropout, Flipout, and Deep Ensembles across calibration, out-of-distribution (OOD) detection. Below, we interpret our main findings, outline their practical implications, and suggest future research.

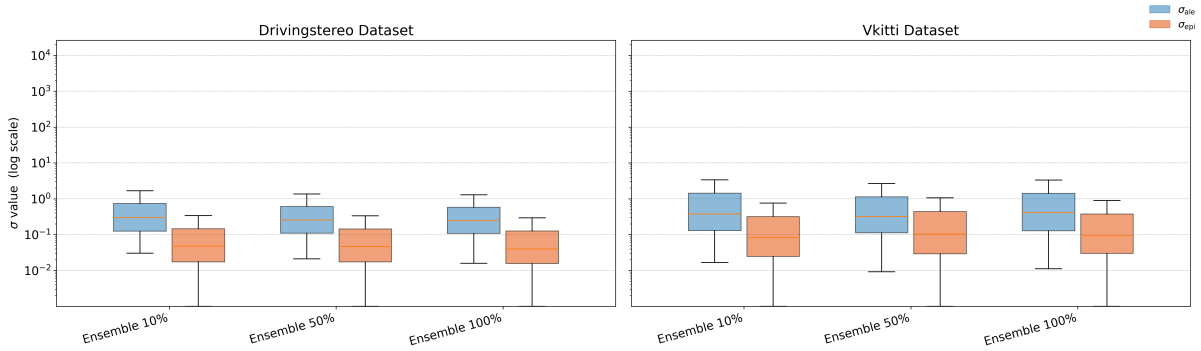
5.1 Key Findings

In-Distribution On the DrivingStereo test set, all methods achieve sub-half-metre MAE , with Deep Ensembles slightly outperforming the others. However, all methods exhibit some degree of overconfidence, as indicated by non-negligible Calibration Errors ($CE \approx 0.32$ for Ensembles, 0.38 for MC Dropout, and 0.41 for Flipout). While ensembles provide relatively better-calibrated uncertainty estimates, none of the approaches perfectly align epistemic uncertainty with actual errors. This suggests a need for post-hoc recalibration techniques to correct miscalibration and improve the reliability of uncertainty estimates.

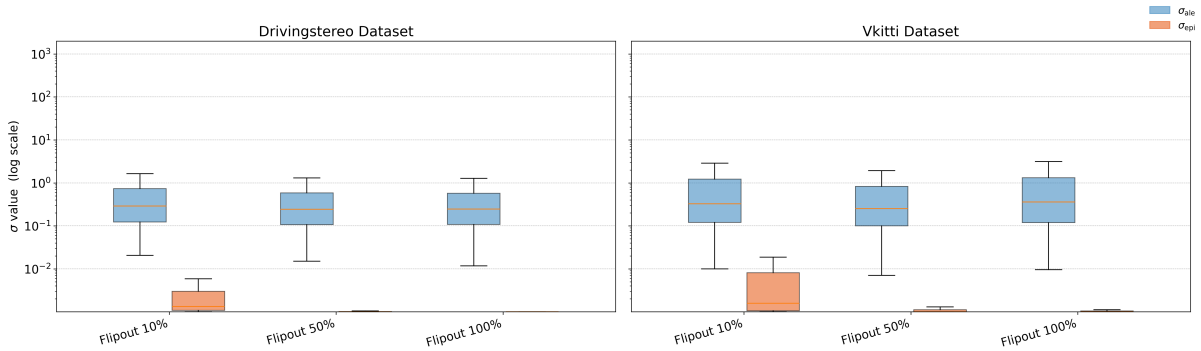
Out-of-Distribution When applied to synthetic vKITTI 2, all methods show slightly increased epistemic uncertainty. Furthermore, in the left–right swap perturbation test, elevated epistemic uncertainty across the entire image for Ensemble and MC Dropout further shows reliable OOD detection. Flipout is shown not to be a reliable OOD detector in this case (as can be seen in Appendix B). However, aleatoric uncertainty remains relatively unchanged between real and synthetic data. Given that synthetic scenes like vKITTI 2 are virtually noise-free, aleatoric uncertainty should be significantly lower. This indicates that current methods fail to adjust aleatoric estimates appropriately under OOD conditions. In the OOD setting (second part of Figure 4.2), predicted depth correlates more strongly with error than any uncertainty estimate.



(a) MC-Dropout



(b) Deep Ensemble

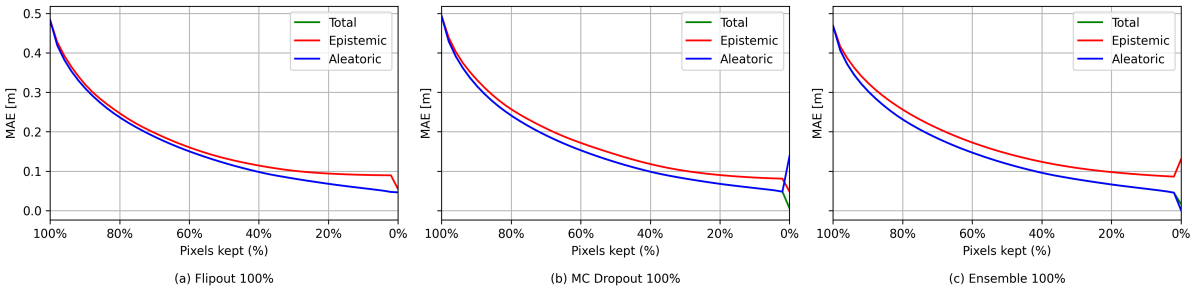


(c) Flipout

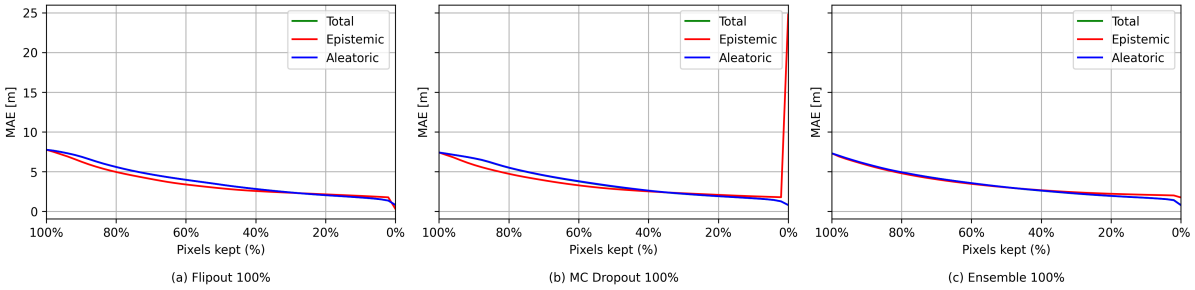
Figure 4.6: Box plots (log-scale on the y–axis) demonstrating distributions of per-pixel AU (blue) and EU (orange) uncertainty for MC Dropout, Flipout, and Deep Ensemble models trained on 10%, 50% and 100%. The DrivingStereo (ID) distributions are on the left and Virtual KITTI 2 (OOD) are on the right. All 3 methods meet the expected uncertainty behaviour: EU decreases with higher training percentages, while AU remains essentially unchanged.

Method	Pearson correlation coefficient r						
	err–σ_{tot}	err–σ_{ale}	err–σ_{epi}	err–$\hat{z}$	σ_{ale}-z_{gt}	σ_{epi}-z_{gt}	σ_{epi}–σ_{ale}
DrivingStereo							
MC Dropout 100%	0.597	0.594	0.525	0.555	0.658	0.556	0.697
Flipout 100%	0.635	**0.635**	0.406	0.567	**0.724**	0.257	**0.588**
Ensemble 100%	0.630	0.617	0.598	0.556	0.695	0.614	0.735
VirtualKITTI2 – OOD							
MC Dropout 100%	0.176	0.174	0.342	0.516	0.260	0.349	0.598
Flipout 100%	0.113	0.113	0.089	**0.523**	0.180	0.064	**0.457**
Ensemble 100%	0.054	0.053	0.312	0.520	0.032	0.276	0.715

Table 4.2: Pearson correlation coefficients r between per-pixel depth error and uncertainty estimates (total, aleatoric, epistemic) and predicted depth ($\hat{z}$), as well as between uncertainty measures and true depth (z_{gt}), for the DrivingStereo (in-distribution) and VirtualKITTI2 (out-of-distribution) datasets. Bold values indicate for each dataset and correlation pairing: for the error columns, the best indicator of depth error (MAE); for the uncertainty–depth columns (DrivingStereo only), the model that best captured noise–depth relationships; and for the epistemic–aleatoric column, the pairing with the weakest (i.e. least) correlation between epistemic and aleatoric uncertainty.



(a) Sparsification on DrivingStereo



(b) Sparsification on Virtual KITTI 2

Figure 4.7: Sparsification curves from MC Dropout, Flipout and Deep Ensemble trained on 100% of the train set: error (MAE) vs. retained fraction for (a) DrivingStereo and (b) Virtual KITTI 2.

This likely arises because the model's predictions span only up to 80 m while the ground-truth extends to 500 m, so the largest predicted depths inherently produce the largest absolute errors. However, in an OOD setting our primary concern is detecting out-of-distribution inputs, not ranking pixels by their error magnitude.

Uncertainty Disentanglement Our analysis of aleatoric and epistemic uncertainty across varying training-set sizes reveals a partial disentanglement but also some degree of contamination. Aleatoric uncertainty remains stable as the training fraction increases from 10% to 100%, reflecting its capture of irreducible observational noise learned even from small subsets of the data, whereas epistemic uncertainty decreases with more data, demonstrating that additional samples help resolve parameter uncertainty.

Table 4.2 quantifies these relationships via Pearson correlation coefficients. On DrivingStereo, the correlation between epistemic and aleatoric uncertainties is high (e.g., $r = 0.735$ for Ensemble at 100% and $r = 0.697$ for Flipout at 100%) and remains positive across methods. This positive relation could indicate that, in practice, aleatoric estimates do not exist in isolation but carry a degree of epistemic uncertainty, as it is also predicted by the model, thereby aligning with the axiomatic claim of Bülte et al. (2025) that predicted variance possesses its own degree of EU.

Flipout producing zero EU In our in-distribution experiments on DrivingStereo, Flipout's estimated epistemic uncertainty collapses to near zero across all training-set fractions. As shown in Fig. 4.6c, the median EU for Flipout remains effectively close to zero for all predicted pixels. Similarly, during our OOD detection protocol, Flipout assigns almost no epistemic uncertainty to OOD inputs, although it is slightly elevated compared to ID. Those findings are consistent with Valdenegro-Toro & Mori (2022), who observed similar behavior on a toy dataset.

The weight distributions learned by Flipout can be very narrow, so even across multiple stochastic forward passes the model exhibits almost no variation—resulting in near-zero estimated epistemic uncertainty.

Practically, this underestimation means Flipout may not be suited for applications requiring reliable uncertainty quantification under model ignorance, such as safety-critical autonomous perception.

5.2 Practical Implications

Method Selection Deep Ensembles offer the best calibration and accuracy, making them suitable when computational resources allow. MC Dropout provides a favorable trade-off, with competitive accuracy and simpler implementation but somewhat poorer calibration. Flipout underestimates epistemic uncertainty, potentially limiting its usibility for safety-critical OOD detection.

Uncertainty-Guided Safety This positive correlation between uncertainty and error demonstrates the potential of these estimates for integration into downstream decision modules—such as triggering fallback planners or engaging sensor fusion when uncertainty crosses a threshold.

5.3 Limitations

Small ensemble size We limited our deep ensemble to $M = 3$ independent models due to computational constraints, which may restrict the diversity of predictions and understate the full potential of ensembles with larger member counts.

Uniform training epochs with varying data fractions All methods were trained for the same number of epochs and with the same batch size. However, because the 10% and 50% subsets comprise only 16,528 and 82,640 stereo pairs respectively—compared to 165,280 pairs in the full set—these experiments involved proportionally fewer weight updates, which may have affected convergence dynamics and the comparability of results across data fractions.

5.4 Future Work

We identify two key avenues for further investigation:

1. **Uncertainty-Guided Active Learning.** Building on the fact that epistemic uncertainty reflects model ignorance and can be reduced by

collecting more data, future work could implement an active learning loop in which pixels or image regions with the highest epistemic uncertainty are selectively annotated and added to the training set. This strategy would exploit EU to maximize information gain per label, potentially achieving comparable depth-estimation performance with far fewer annotations.

2. **Post-hoc Recalibration of Uncertainty Estimates.** While our models exhibit miscalibration, techniques such as temperature scaling or isotonic regression can be applied after training to adjust predicted variances so that nominal confidence levels better match empirical error frequencies. Evaluating these post-hoc calibration methods on stereo-depth predictors may yield more trustworthy uncertainty scores without the need for costly retraining.

3. **Temporal Consistency.** Future work should leverage the continual feedback an autonomous vehicle receives by incorporating temporal cues into stereo-depth prediction. Integrating motion flow, temporal ensembling, or lightweight recurrent updates could smooth transient errors and detect abrupt uncertainty spikes, enhancing real-time depth accuracy and reliability.

6 Conclusion

In this thesis, we set out to answer three core research questions concerning uncertainty quantification methods applied to PSMNet in stereo depth estimation.

1. Calibration We evaluated calibration via reliability diagrams and the Calibration Error (CE). Deep Ensembles achieved the lowest CE on both ID and OOD data, MC Dropout ranked second, and Flipout (whose epistemic uncertainty collapsed to near zero) was last. None of the methods can be considered well calibrated. All three require post-hoc calibration to align their nominal confidence intervals with observed error coverage, especially in safety-critical applications.

2. Disentanglement By training on 10%, 50%, and 100% of DrivingStereo, we observed that aleatoric uncertainty remains essentially invariant to dataset size—capturing irreducible sensor noise—whereas epistemic uncertainty decreases as more data is provided, demonstrating its sensitivity to model ignorance. However, some positive correlation between the two uncertainty types indicates partial contamination.

3. Out-of-Distribution Detection We tested on synthetic Virtual KITTI 2 and on left–right image swaps. On Virtual KITTI 2, both Deep Ensembles and MC Dropout produced elevated epistemic uncertainty for OOD scenes, whereas Flipout failed to register any increase in epistemic variance.

For the left–right swap test, where we feed a left image as the right view and vice versa (introducing physically implausible geometry), Deep Ensembles and MC Dropout again showed marked spikes in epistemic uncertainty over regions with inconsistent depth, validating their ability to detect domain violations. Flipout, in contrast, remained overconfident, underestimating its uncertainty on these corrupted inputs. Aleatoric uncertainty in all methods stayed at levels similar to in-distribution, and thus does not reliably signal domain shift.

In summary, by evaluating and contrasting three uncertainty quantification methods, this thesis advances our understanding of how to build more reliable and safe depth-estimation pipelines for real-world autonomous systems.

References

Bülte, C., Sale, Y., Löhr, T., Hofman, P., Kutyniok, G., & Hüllermeier, E. (2025). An axiomatic assessment of entropy-and variance-based uncertainty quantification in regression. *arXiv preprint arXiv:2504.18433*.

Cabon, Y., Murray, N., & Humenberger, M. (2020). Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.

Chang, J.-R., & Chen, Y.-S. (2018). Pyramid stereo matching network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5410–5418).

de Jong, I. P., Sburlea, A. I., & Valdenegro-Toro, M. (2024). How disentangled are your classification uncertainties? *arXiv preprint arXiv:2408.12175*.

Du, H., Li, Y., Sun, Y., Zhu, J., & Tombari, F. (2021). Srh-net: Stacked recurrent hourglass network for stereo matching. *IEEE Robotics and Automation Letters*, *6*(4), 8005–8012.

Eser, A. Y. (2018, 11). *Derinlik i: Stereo kalibrasyon ve düzeltme - İyi programlama - medium.* İyi Programlama. Retrieved 2025-07-27, from https://medium.com/i%CC%87yi-programlama/derinlik-i-stereo-kalibrasyon-ve-d%C3%BCzeltme-e069c5941469

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Gal, Y., Hron, J., & Kendall, A. (2017). Concrete dropout. *Advances in neural information processing systems*, *30*.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 ieee conference on computer vision and pattern recognition* (pp. 3354–3361).

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision.* Cambridge university press.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, *110*(3), 457–506.

Kanbara, M., Fujii, H., Takemura, H., & Yokoya, N. (2000). A stereo vision-based augmented reality system with an inertial sensor. In *Proceedings ieee and acm international symposium on augmented reality (isar 2000)* (pp. 97–100).

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, *30*.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., & Bry, A. (2017). End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the ieee international conference on computer vision* (pp. 66–75).

Laga, H., Jospin, L. V., Boussaid, F., & Bennamoun, M. (2020). A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, *44*(4), 1738–1764.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

Mehltretter, M. (2022). Joint estimation of depth and its uncertainty from stereo images using bayesian deep learning. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*, 69–78.

Poggi, M., Aleotti, F., Tosi, F., & Mattoccia, S. (2020). On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3227–3237).

Probst, T., Maninis, K.-K., Chhatkuli, A., Ourak, M., Vander Poorten, E., & Van Gool, L. (2017). Automatic tool landmark detection for stereo vision in robot-assisted retinal surgery. *IEEE Robotics and Automation Letters*, *3*(1), 612–619.

Rao, Z., Xiong, B., He, M., Dai, Y., He, R., Shen, Z., & Li, X. (2023). Masked representation learning for domain generalized stereo matching. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5435–5444).

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, *47*(1), 7–42.

Valdenegro-Toro, M., & Mori, D. S. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (pp. 1508–1516).

Wang, C., Wang, X., Zhang, J., Zhang, L., Bai, X., Ning, X., ... Hancock, E. (2022). Uncertainty estimation for stereo matching based on evidential deep learning. *pattern recognition*, *124*, 108498.

Wen, Y., Vicol, P., Ba, J., Tran, D., & Grosse, R. (2018). Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., & Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence* (pp. 2282–2292).

Xu, G., Cheng, J., Guo, P., & Yang, X. (2022). Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12981–12990).

Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., & Zhou, B. (2019). Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 899–908).

A Additional Visualizations

A.1 Qualitative Comparison

In this section, we qualitatively compare the depth maps, aleatoric and epistemic uncertainty maps produced by MC Dropout, Ensembles, and Flipout. We highlight both the shared behaviors and the key differences in how strongly and where each approach expresses its uncertainty over predictions.

Depth Maps As can be seen in Figure A.1 (Ensemble), Figure A.2 (MC Dropout) and Figure A.3 (Flipout), all three methods exhibit a tendency to hallucinate spurious depth in the uppermost regions of the scene, likely due to the sparsity of the DrivingStereo dataset, where valid depth values are often absent in those areas. Flipout seems to hallucinate the least on this ID sample. However, this behavior does not hold when evaluated on the out-of-distribution Virtual KITTI 2 dataset. As shown in Figure A.6, Flipout actually hallucinates more non-existent objects (particularly in the upper regions of the image) than either the deep ensemble (Figure A.4) or MC Dropout (Figure A.5). All methods predict the synthetic rain drops to be objects in the scene.

Aleatoric Uncertainty Maps Although the overall spatial patterns of aleatoric uncertainty are consistent across all three methods, the relative intensity differs by region. Flipout concentrates higher aleatoric responses along true object edges compared to the other two methods (AU maps in Figures A.3 and A.6), MC Dropout intensifies its AU uncertainty levels around the boundaries of its hallucinated structures (Figures A.2 and A.5), and the deep ensemble distributes its aleatoric uncertainties more evenly between real and spurious boundaries (Figures A.1 and A.4).

Epistemic Uncertainty Maps The epistemic maps exhibit the greatest variability across methods. Flipout produces minimal EU level in both in-distribution and out-of-distribution cases (EU maps in Figures A.3 and A.6). MC Dropout's epistemic map closely mirrors its aleatoric pattern, with heightened uncertainty along the edges of both real and hallucinated objects (Figures A.2 and A.5). In contrast, the deep ensemble displays its strongest epistemic response within of the hallucinated objects (Figures A.1 and A.4), as detailed in Section 4.

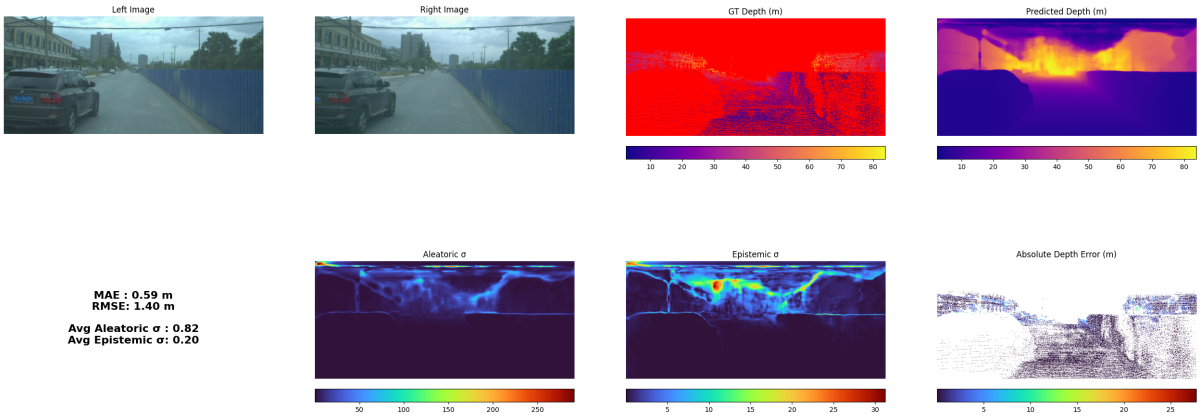


Figure A.1: Ensemble 100%: Qualitative results (DrivingStereo) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 0.59 m, RMSE = 1.40 m, average AU = 0.82 and average EU = 0.20.

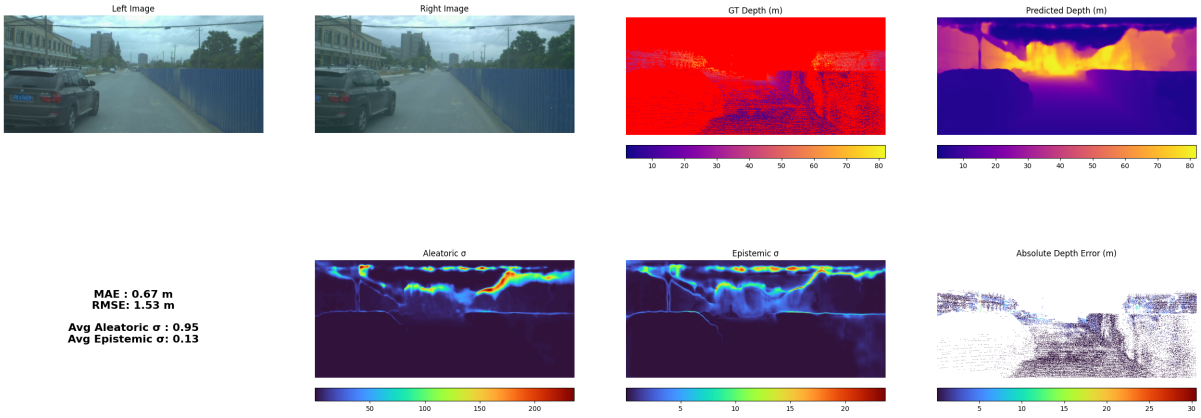


Figure A.2: MC Dropout 100%: Qualitative results (DrivingStereo) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 0.67 m, RMSE = 1.53 m, average AU = 0.95 and average EU = 0.13.

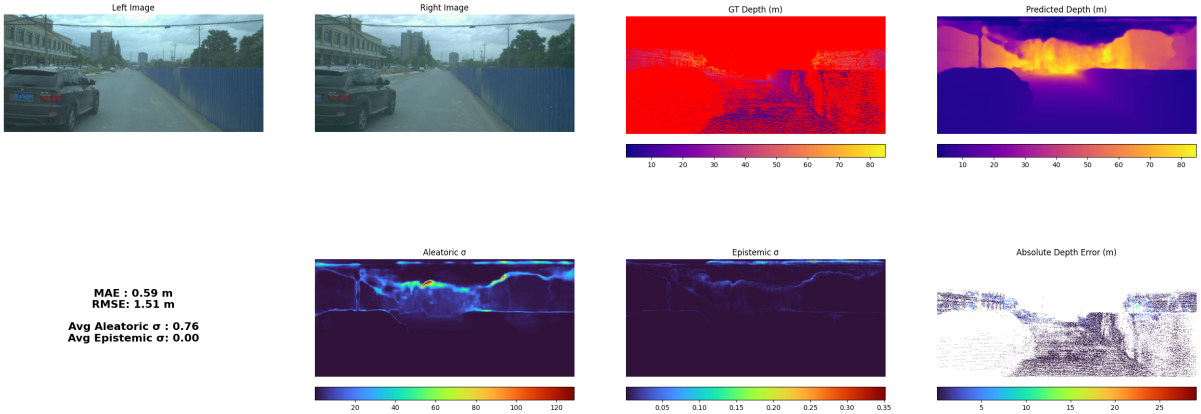


Figure A.3: Flipout 100%: Qualitative results (DrivingStereo) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 0.59 m, RMSE = 1.51 m, average AU = 0.76 and average EU = 0.00.

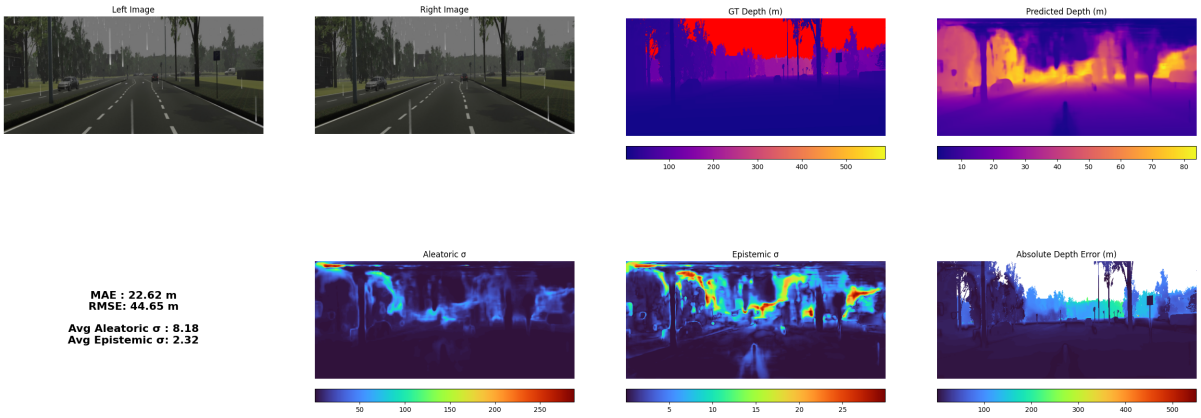


Figure A.4: Ensemble 100%: Qualitative results (Virtual Kitti 2) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 22.62 m, RMSE = 44.65 m, average AU = 8.18 and average EU = 2.32.

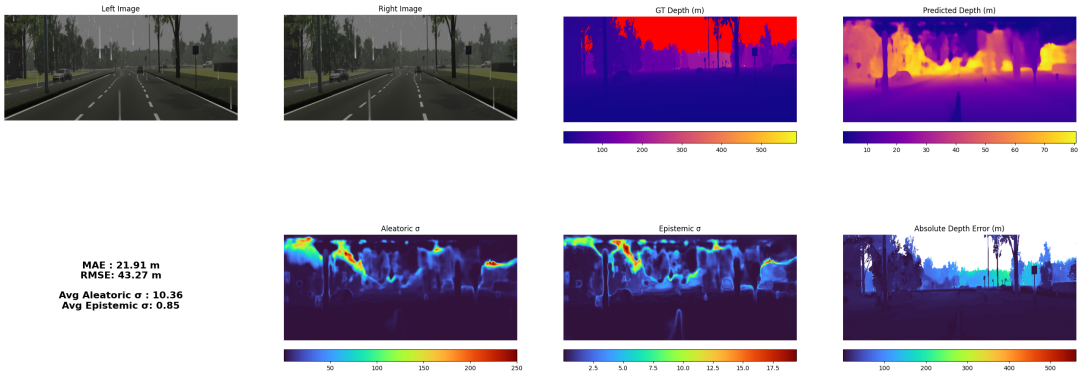


Figure A.5: MC Dropout 100%: Qualitative results (Virtual Kitti 2) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 21.91 m, RMSE = 43.27 m, average AU = 10.36 and average EU = 0.85.

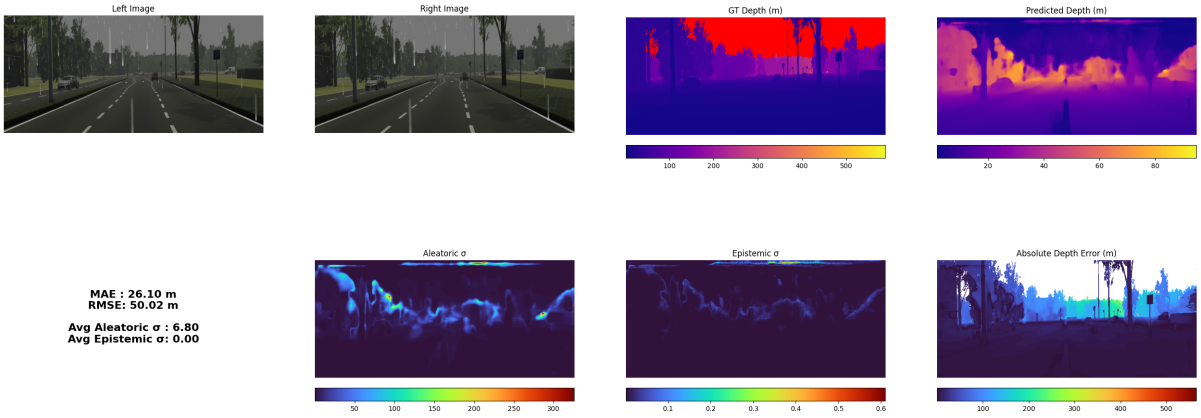


Figure A.6: Flipout 100%: Qualitative results (Virtual Kitti 2) demonstrating the left and right input images, the ground-truth depth map (invalid pixels in red), the predicted depth map, the aleatoric uncertainty map, the epistemic uncertainty map, and the absolute depth error map (invalid pixels in white), with MAE = 26.10 m, RMSE = 50.02 m, average AU = 6.80 and average EU = 0.00.

B Left–right image swap tests on MC Dropout and Flipout

In this section we present the left-right image swap test, where we swap the left and right images before feeding them to the network and compare the resulting EU maps, as they should indicate an OOD sample. We expect that swapping the inputs will break the network's ability to infer correct disparity—producing a highly corrupted depth map—and that this failure will be accompanied by a rise in EU.

MC Dropout As illustrated by the depth estimates in Figure B.1, feeding the models swapped left–right inputs yields incoherent depth reconstructions. Across the entire scene, the epistemic uncertainty is markedly elevated relative to the correctly ordered input, demonstrating effective out-of-distribution detection.

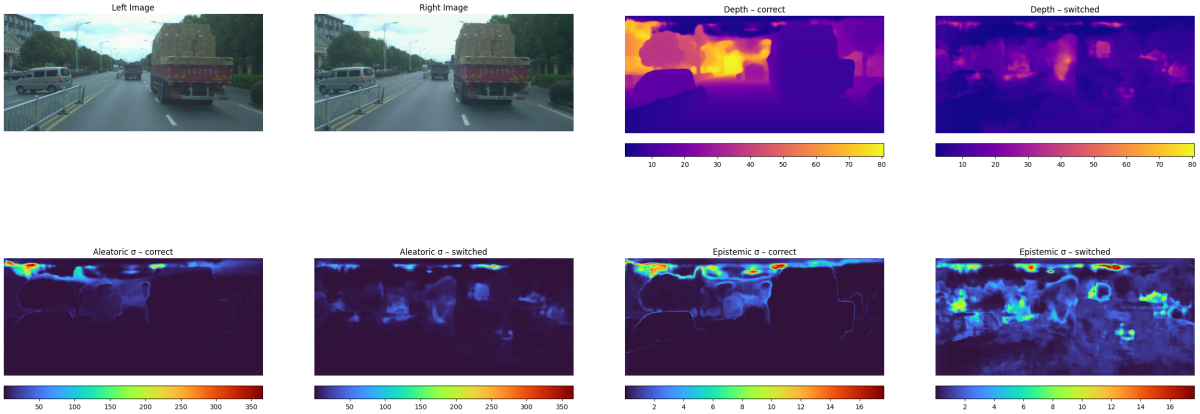


Figure B.1: MC Dropout 100%: Qualitative Results (DrivingStereo) from left-right input swap test. Panels labeled "correct" show properly paired stereo inputs, while panels labeled "switched" were obtained by swapping the left and right inputs. As desired, epistemic uncertainty (EU) increases in the switched scenario.

Flipout Like the other methods, Flipout produces a poor depth map when its inputs are swapped (Figure B.2). However, unlike the others, it exhibits only a minimal increase in epistemic uncertainty compared to the correctly ordered input, indicating weak OOD detection.

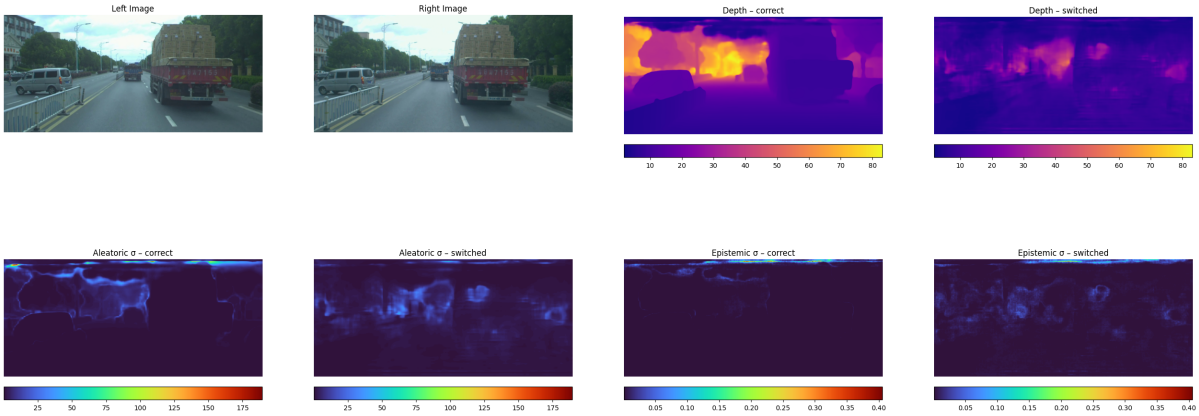


Figure B.2: Flipout 100%: Qualitative Results (DrivingStereo) from left-right input swap test. Panels labeled "correct" show properly paired stereo inputs, while panels labeled "switched" were obtained by swapping the left and right inputs. EU exhibits little to no increase in the switched scenario relative to the correctly paired inputs, indicating Flipout's poor OOD detection.

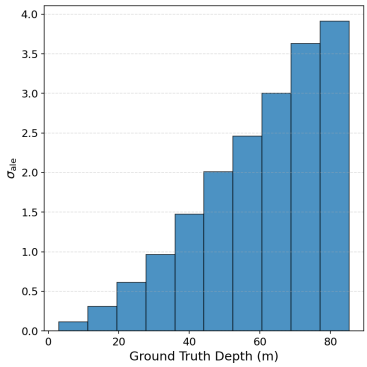
C Supplementary Plots

This appendix collects bar plots figures for the three uncertainty estimation methods: Ensemble, Flipout, and MC Dropout, all trained on 100% of the train set. All bar-chart panels in this appendix plot a summary metric (AU or EU) against either depth or error by partitioning the x-axis into ten uniform-width bins and displaying the mean metric value within each bin. Because the bins are evenly spaced but sample counts can vary significantly depending on the distribution of values along the axis, some bins contain many more samples than others, which can influence the stability and representativeness of the mean estimates.

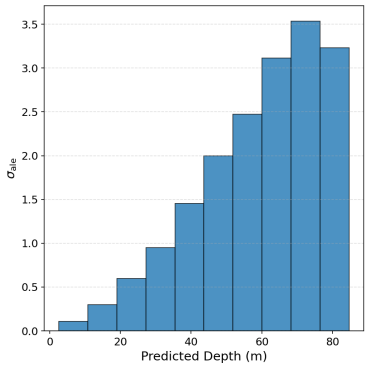
Ensemble Figure C.1a and C.1b show that higher aleatoric uncertainty corresponds to larger mean depths in both the ground-truth and predicted bins. A very similar trend holds for epistemic uncertainty (Figures C.1c and C.1d). As epistemic uncertainty increases, the mean absolute error (MAE) within each bin also rises (see Figure C.1f). In contrast, as can be seen in Figure C.1e, aleatoric uncertainty does not exhibit such a clear relationship with MAE. A positive relationship is observed between MAE and ground-truth depth (Figure C.1g), as well as between MAE and predicted depth (Figure C.1h).

MC Dropout Figure C.2 presents the bar plots for the MC Dropout method. As with Ensembles, higher aleatoric uncertainty is associated with larger mean depths in both the ground-truth and predicted depth bins (see Figure C.2a and C.2b. Epistemic uncertainty shows a similar pattern (Figure C.2c), although the increase in predicted depth is less linear (Figure C.2d). As observed in Figures C.2e and C.2f, neither aleatoric nor epistemic uncertainty consistently rises with MAE. On the other hand, ground-truth depth and predicted depth both have a positive relation with MAE (Figures C.2g and C.2h).

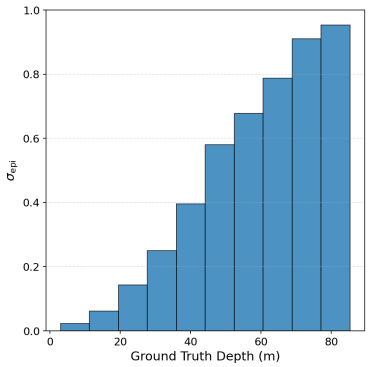
Flipout Figures C.3a and C.3b) illustrate that, for Flipout, increasing aleatoric uncertainty coincides with higher mean depths in both the ground-truth and predicted bins. Epistemic uncertainty remains relatively flat across most depth ranges, only rising noticeably beyond 60 m (Figures C.3c and C.3d). Moreover, aleatoric uncertainty correlates positively with MAE—bins where greater AU exhibit larger errors (Figure C.3e), whereas epistemic uncertainty displays the opposite trend (see Figure C.3f). Similarly to the other 2 methods, MAE has a positive relation with ground-truth and predicted depth (Figures C.3g and C.3h).



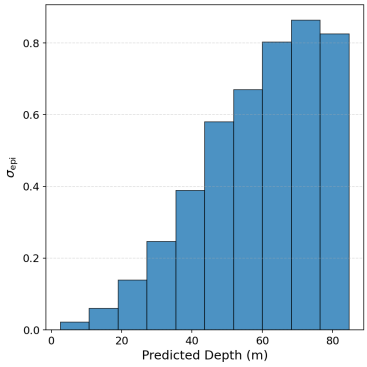
(a) σ_{ale} vs. ground-truth depth.



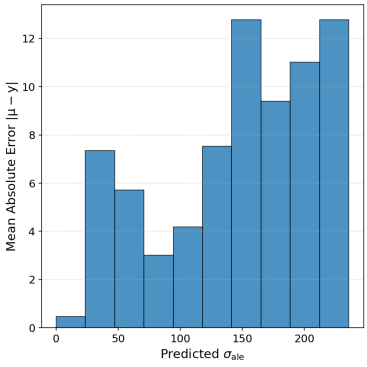
(b) σ_{ale} vs. predicted depth.



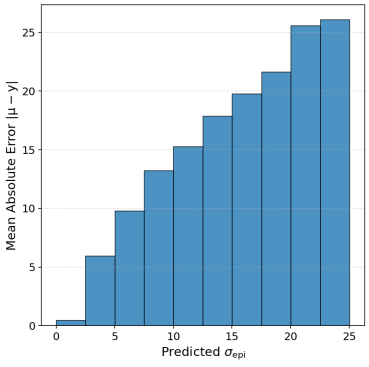
(c) σ_{epi} vs. ground-truth depth.



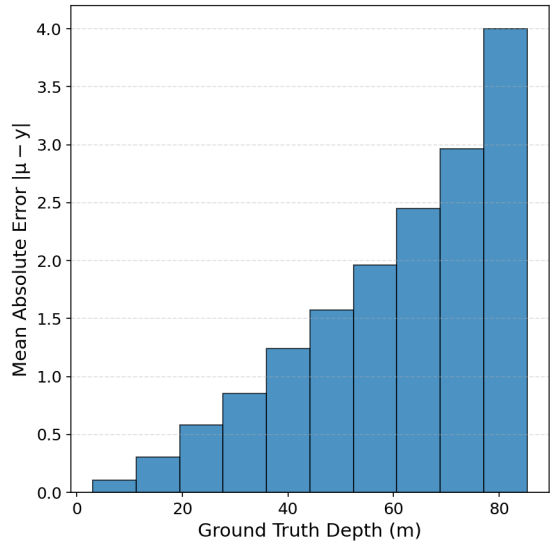
(d) σ_{epi} vs. predicted depth.



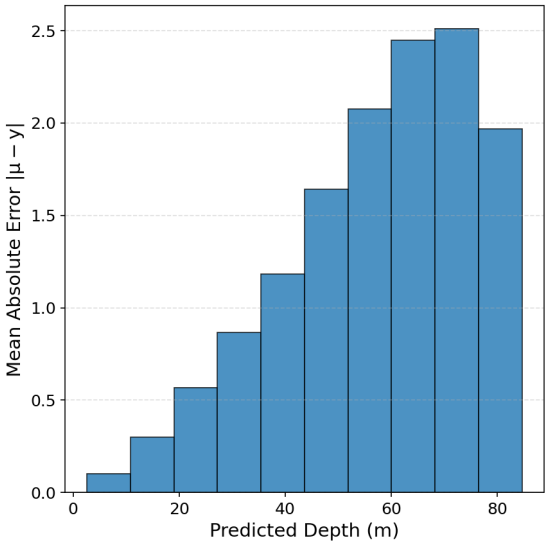
(e) MAE vs. σ_{ale}.



(f) MAE vs. σ_{epi}.

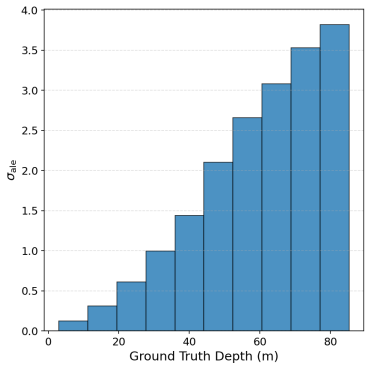


(g) MAE vs. ground-truth depth.

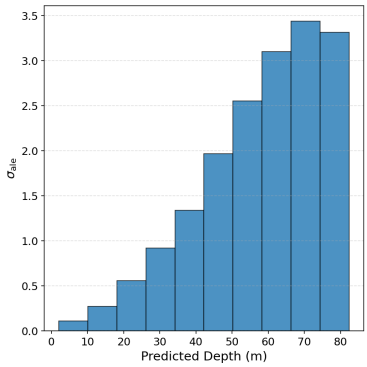


(h) MAE vs. predicted depth.

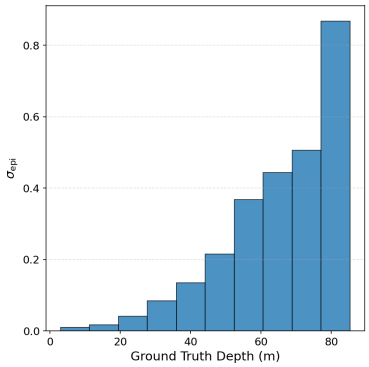
Figure C.1: Bar charts with 10 uniform bins for the Ensemble method on the DrivingStereo dataset (100%) designed to test the relationships between depth, uncertainty (AU and EU), and error (MAE). In particular, panel (a) shows the strongest positive association with ground-truth depth among the uncertainties, and panels (f) and (g) show the strongest positive relation with MAE.



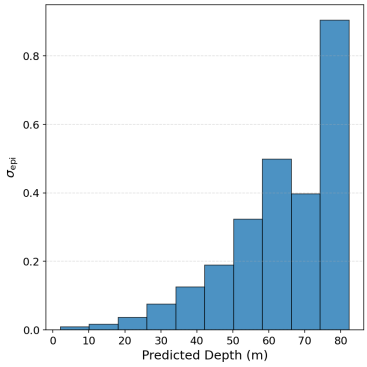
(a) σ_{ale} vs. ground-truth depth.



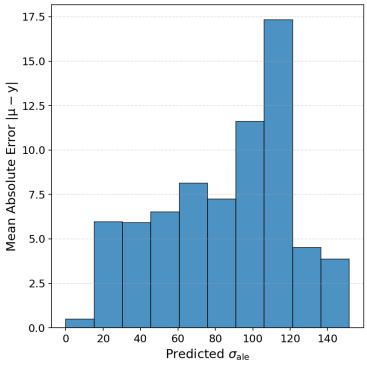
(b) σ_{ale} vs. predicted depth.



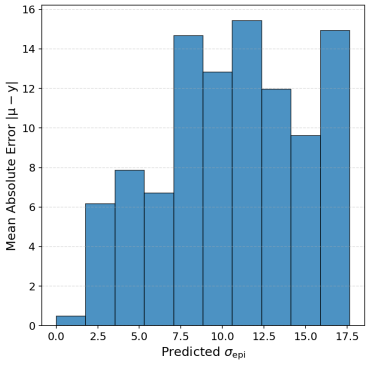
(c) σ_{epi} vs. ground-truth depth.



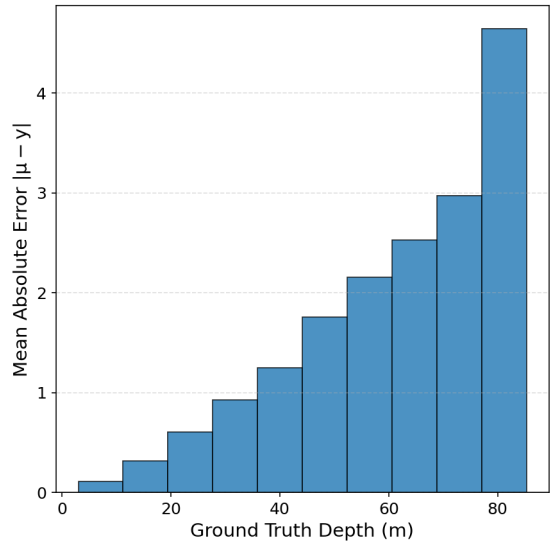
(d) σ_{epi} vs. predicted depth.



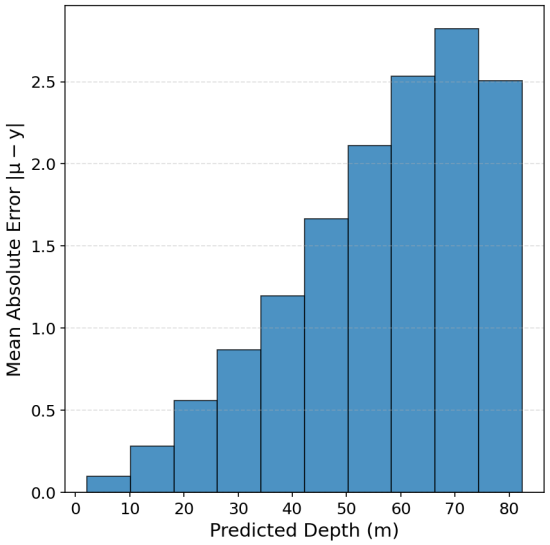
(e) MAE vs. σ_{ale}.



(f) MAE vs. σ_{epi}.

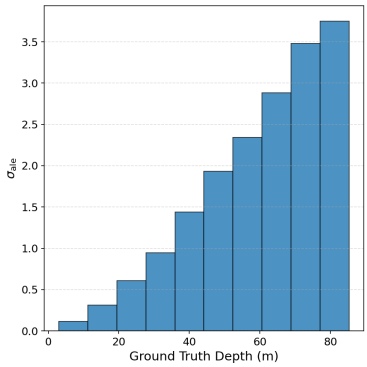


(g) MAE vs. ground-truth depth.

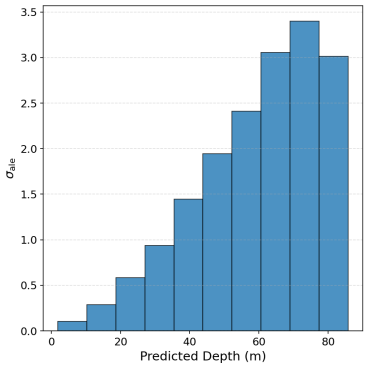


(h) MAE vs. predicted depth.

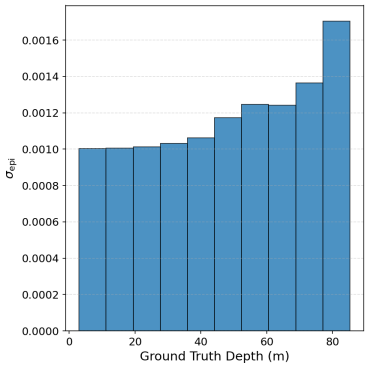
Figure C.2: Bar charts with 10 uniform bins for MC Dropout method on the DrivingStereo dataset (100%) designed to test the relationships between depth, uncertainty (AU and EU), and error (MAE). In particular, panels (a) and (c) show the strongest positive association with ground-truth depth among the uncertainties, and panel (g) show the strongest positive relation with MAE.



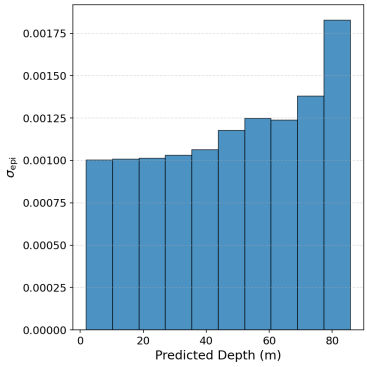
(a) σ_{ale} vs. ground-truth depth.



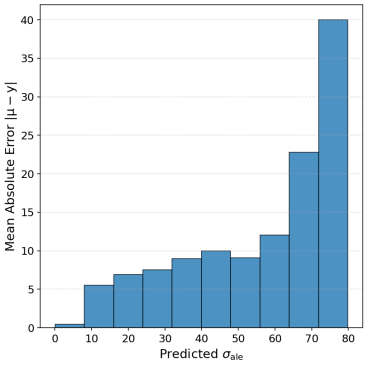
(b) σ_{ale} vs. predicted depth.



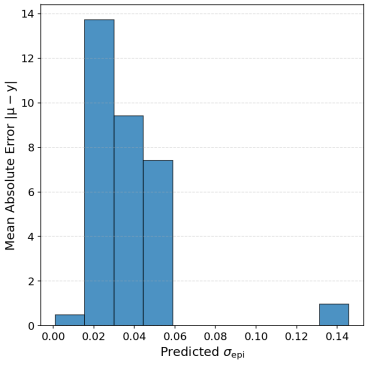
(c) σ_{epi} vs. ground-truth depth.



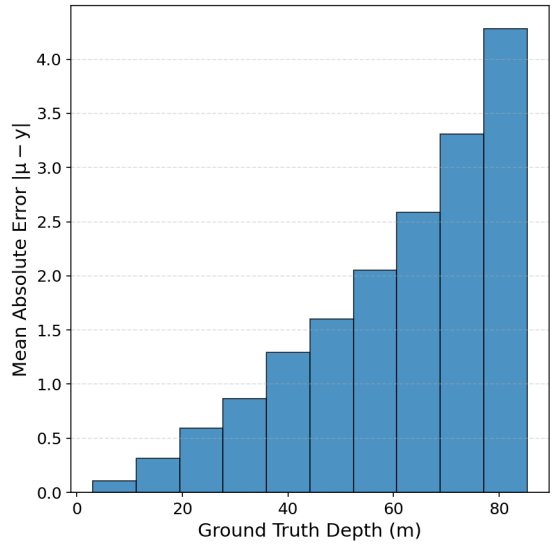
(d) σ_{epi} vs. predicted depth.



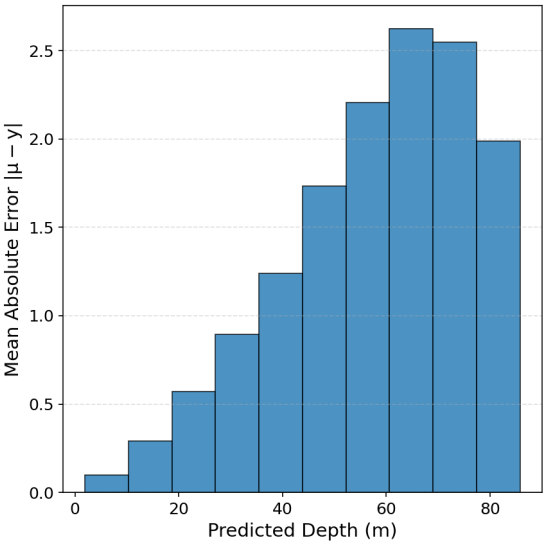
(e) MAE vs. σ_{ale}.



(f) MAE vs. σ_{epi}.



(g) MAE vs. ground-truth depth.



(h) MAE vs. predicted depth.

Figure C.3: Bar charts with 10 uniform bins for Flipout method on the DrivingStereo dataset (100%) designed to test the relationships between depth, uncertainty (AU and EU), and error (MAE). In particular, panel (a) shows the strongest positive association with ground-truth depth among the uncertainties, and panel (g) show the strongest positive relation with MAE.

D Hyperparameter Settings

D.1 General Training Parameters

This applies to every method and each training size:

- **Maximum Disparity:** 224
- **Number of Epochs:** 30
- **Batch Size:** 8
- **Learning Rate:** 1×10^{-3}

D.2 Model-Specific Parameters

Dropout

- **Dropout Type:** Monte Carlo Dropout (MC Dropout)
- **Dropout Rate:** 0.3

Flipout Model

- **Bayesian Layer:** `Conv3dFlipout` from `bayesian-torch`
- **Prior Mean:** 0.0
- **Prior Variance:** 1.0
- **KL Accumulation:** KL divergence term accumulated from the Bayesian layer

D.3 Pre-processing

- **Image normalization:** Left and right RGB images were normalized using ImageNet statistics with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].
- **Resizing:** All images and disparity maps were resized to a uniform resolution of 672×304 pixels.