Supplementary Materials

S1 Derivations

S1.1 Normal Inverse-Gamma moments

We assume our data was drawn from a Gaussian with unknown mean and variance, (μ, σ^2). We probabilistically model these parameters, θ, according to:

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}) \tag{S1}$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta). \tag{S2}$$

Therefore, the prior joint distribution can be written as:

$$p(\underbrace{\mu, \sigma^2}_{\boldsymbol{\theta}} \,|\, \underbrace{\gamma, \upsilon, \alpha, \beta}_{\boldsymbol{m}}) = p(\mu)\, p(\sigma^2) \tag{S3}$$
$$= \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1})\, \Gamma^{-1}(\alpha, \beta) \tag{S4}$$
$$= \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2}\right\}. \tag{S5}$$

The first order moments of this distribution represent the maximum likelihood prediction as well as uncertainty (both aleatoric and epistemic).

$$\mathbb{E}[\mu] = \int_{\mu=-\infty}^{\infty} \mu\, p(\mu)\, \mathrm{d}\mu = \gamma \tag{S6}$$

$$\mathbb{E}[\sigma^2] = \int_{\sigma^2=0}^{\infty} \sigma^2\, p(\sigma^2)\, \mathrm{d}\sigma^2 \tag{S7}$$
$$= \int_{\sigma=0}^{\infty} \sigma^2\, p(\sigma^2)\, (2\sigma)\, \mathrm{d}\sigma \tag{S8}$$
$$= \frac{\beta}{\alpha - 1}, \qquad \forall\, \alpha > 1 \tag{S9}$$

$$\mathrm{Var}[\mu] = \int_{\mu=-\infty}^{\infty} \mu^2\, p(\mu)\, \mathrm{d}\mu - (\mathbb{E}[\mu])^2 \tag{S10}$$
$$= \gamma^2 - \frac{\sigma^2}{\upsilon} - (\mathbb{E}[\mu])^2 \tag{S11}$$
$$= \gamma^2 - \frac{\frac{\beta}{\alpha-1}}{\upsilon} - \gamma^2 \tag{S12}$$
$$= \frac{\beta}{\upsilon(\alpha-1)}, \qquad \forall\, \alpha > 1 \tag{S13}$$

In summary,

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \qquad \underbrace{\mathbb{E}[\sigma^2] = \tfrac{\beta}{\alpha-1}}_{\text{aleatoric}}, \qquad \underbrace{\mathrm{Var}[\mu] = \tfrac{\beta}{\upsilon(\alpha-1)}}_{\text{epistemic}}. \tag{S14}$$

S1.2 Model evidence & Type II Maximum Likelihood Loss

In this subsection, we derive the posterior predictive or model evidence (ie. Eq. 7) of a NIG distribution. Marginalizing out μ and σ gives our desired result:

$$p(y_i|\boldsymbol{m}) = \int_{\boldsymbol{\theta}} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) \, \mathrm{d}\boldsymbol{\theta} \tag{S15}$$

$$= \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} p(y_i|\mu,\sigma^2) p(\mu,\sigma^2|\boldsymbol{m}) \, \mathrm{d}\mu \, \mathrm{d}\sigma^2 \tag{S16}$$

$$= \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} p(y_i|\mu,\sigma^2) p(\mu,\sigma^2|\gamma,\upsilon,\alpha,\beta) \, \mathrm{d}\mu \, \mathrm{d}\sigma^2 \tag{S17}$$

$$= \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} \left[\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i-\mu)^2}{2\sigma^2} \right\} \right] \tag{S18}$$

$$\left[\frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{ -\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2} \right\} \right] \mathrm{d}\mu \, \mathrm{d}\sigma^2 \tag{S19}$$

$$= \int_{\sigma^2=0}^{\infty} \frac{\beta^\alpha \sigma^{-3-2\alpha}}{\sqrt{2\pi}\sqrt{1+1/\upsilon}\Gamma(\alpha)} \exp\left\{ -\frac{2\beta + \frac{\upsilon(y_i-\gamma)^2}{1+\upsilon}}{2\sigma^2} \right\} \mathrm{d}\sigma^2 \tag{S20}$$

$$= \int_{\sigma=0}^{\infty} \frac{\beta^\alpha \sigma^{-3-2\alpha}}{\sqrt{2\pi}\sqrt{1+1/\upsilon}\Gamma(\alpha)} \exp\left\{ -\frac{2\beta + \frac{\upsilon(y_i-\gamma)^2}{1+\upsilon}}{2\sigma^2} \right\} 2\sigma \, \mathrm{d}\sigma \tag{S21}$$

$$= \frac{\Gamma(1/2+\alpha)}{\Gamma(\alpha)} \sqrt{\frac{\upsilon}{\pi}} \left(2\beta(1+\upsilon)\right)^\alpha \left(\upsilon(y_i-\gamma)^2 + 2\beta(1+\upsilon)\right)^{-(\frac{1}{2}+\alpha)} \tag{S22}$$

$$p(y_i|\boldsymbol{m}) = \mathrm{St}\left(y_i; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}, 2\alpha\right). \tag{S23}$$

$\mathrm{St}\left(y; \mu_{\mathrm{St}}, \sigma^2_{\mathrm{St}}, \upsilon_{St}\right)$ is the Student-t distribution evaluated at y with location parameter μ_{St}, scale parameter σ^2_{St}, and υ_{St} degrees of freedom. Using this result we can compute the negative log likelihood loss, $\mathcal{L}_i^{\mathrm{NLL}}$, for sample i as:

$$\mathcal{L}_i^{\mathrm{NLL}} = -\log p(y_i|\boldsymbol{m}) \tag{S24}$$

$$= -\log\left(\mathrm{St}\left(y_i; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}, 2\alpha\right)\right) \tag{S25}$$

$$\mathcal{L}_i^{\mathrm{NLL}} = \tfrac{1}{2}\log\left(\tfrac{\pi}{\upsilon}\right) - \alpha\log(\Omega) + \left(\alpha+\tfrac{1}{2}\right)\log((y-\gamma)^2\upsilon + \Omega) + \log\left(\tfrac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right) \tag{S26}$$

where $\Omega = 2\beta(1+\upsilon)$.

S1.3 KL-divergence of the Normal Inverse-Gamma

The KL-divergence between two Normal Inverse-Gamma functions is given by [44]:

$$\mathbb{KL}\big(p(\mu,\sigma^2|\gamma_1,\upsilon_1,\alpha_1,\beta_1) \,||\, p(\mu,\sigma^2|\gamma_2,\upsilon_2,\alpha_2,\beta_2)\big) \tag{S27}$$

$$= \mathbb{KL}\big(\texttt{NIG}(\gamma_1,\upsilon_1,\alpha_1,\beta_1) \,||\, \texttt{NIG}(\gamma_2,\upsilon_2,\alpha_2,\beta_2)\big) \tag{S28}$$

$$= \frac{1}{2}\frac{\alpha_1}{\beta_1}(\mu_1-\mu_2)^2\upsilon_2 + \frac{1}{2}\frac{\upsilon_2}{\upsilon_1} - \frac{1}{2} + \alpha_2\log\left(\frac{\beta_1}{\beta_2}\right) - \log\left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)}\right) \tag{S29}$$

$$+ (\alpha_1-\alpha_2)\Psi(\alpha_1) - (\beta_1-\beta_2)\frac{\alpha_1}{\beta_1} \tag{S30}$$

$\Gamma(\cdot)$ is the Gamma function and $\Psi(\cdot)$ is the Digamma function. For zero evidence, both $\alpha = 0$ and $\upsilon = 0$. To compute the KL divergence between one NIG distribution and another with zero evidence we can set either $\upsilon_2 = \alpha_2 = 0$ (i.e., reverse-KL) in which case, $\Gamma(0)$ is not well defined, or

$\upsilon_1 = \alpha_1 = 0$ (i.e. forward-KL) which causes a divide-by-zero error of υ_1. In either approach, the KL-divergence between an arbitrary NIG and one with zero evidence cannot be evaluated.

Instead, we briefly consider a naive alternative which can be obtained by considering an ϵ amount of evidence, where ϵ is a small constant (instead of strictly 0-evidence). This approach yields a well-defined KL-divergence (with fixed γ, β at the consequence of a hyper-sensitive ϵ parameter.

$$\mathbb{KL}\big(\texttt{NIG}(\gamma, \upsilon, \alpha, \beta) \,||\, \texttt{NIG}(\gamma, \epsilon, 1+\epsilon, \beta)\big) \tag{S31}$$

$$= \frac{1}{2}\frac{1+\epsilon}{\upsilon} - \frac{1}{2} - \log\left(\frac{\Gamma(\alpha)}{\Gamma(1+\epsilon)}\right) + (\alpha - (1+\epsilon))\Psi(\alpha) \tag{S32}$$

In Fig. S1.3 we compare the performance of the KL-divergence regularizer compared to our more direct evidence regularizer, for several realizations of the regularization coefficient, λ. We observed extreme sensitivity to the setting of ϵ for different datasets such that we could not achieve the desired regularizing effect for any regularization amount, λ. Unless otherwise stated, all results were obtained using our direct evidence regularizer instead (Eq. 9).

S2 Benchmark regression tasks

S2.1 Cubic toy examples

S2.1.1 Dataset and experimental setup

The training set consists of training examples drawn from $y = x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 3)$ in the region $-4 \leq x \leq 4$, whereas the test data is unbounded (we show in the region $-6 \leq x \leq 6$). This problem setup is identical to that presented in [20, 28]. All models consisted of 100 neurons with 3 hidden layers and were trained to convergence. The data presented in Fig. S1 illustrates the estimated epistemic uncertainty and predicted mean accross the entire test set. Sampling based models [5, 9, 28] used $n = 5$ samples. The evidential model used $\lambda = 0.01$. All models were trained with the Adam optimizer η =5e-3 for 5000 iterations and a batch size of 128.

S2.1.2 Baselines

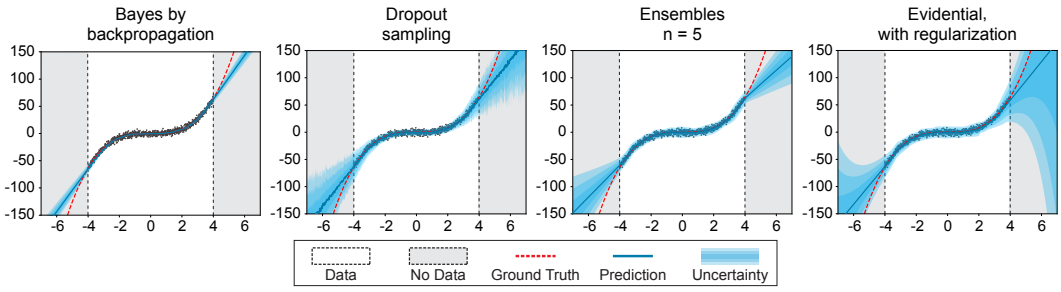

Figure S1: **Epistemic uncertainty estimation baselines** on the dataset $y = x^3 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 3)$.

S2.1.3 Impact of the evidential regularizer

In the following experiment, we demonstrate the importance of augmenting the training objective with our evidential regularizer $\mathcal{L}^R$ as introduced in Sec. 3.3. Fig. S2 provides quantitative results on epistemic uncertainty estimation after training on the same regression problem presented in S2.1 with different realizations of the regularization coefficients, λ. We show the performance of our ability to calibrate uncertainty on OOD data is heavily related to our regularizer. As we decrease our regularizer weight, uncertainty on OOD examples decays to zero. Stronger regularization inflates the uncertainty ($\lambda = 0.01$ is a good choice for this problem) while aleatoric uncertainty is maintained constant. Please refer to Fig. 3 for the regularization effect on both aleatoric and epistemic uncertainty.

S2.1.4 Disentanglement of aleatoric and epistemic uncertainty

In the following experiment, we provide results to suggest that the evidential regularizer is capable of disentangling aleatoric and epistemic uncertainties by capturing incorrect evidence. Specifically,

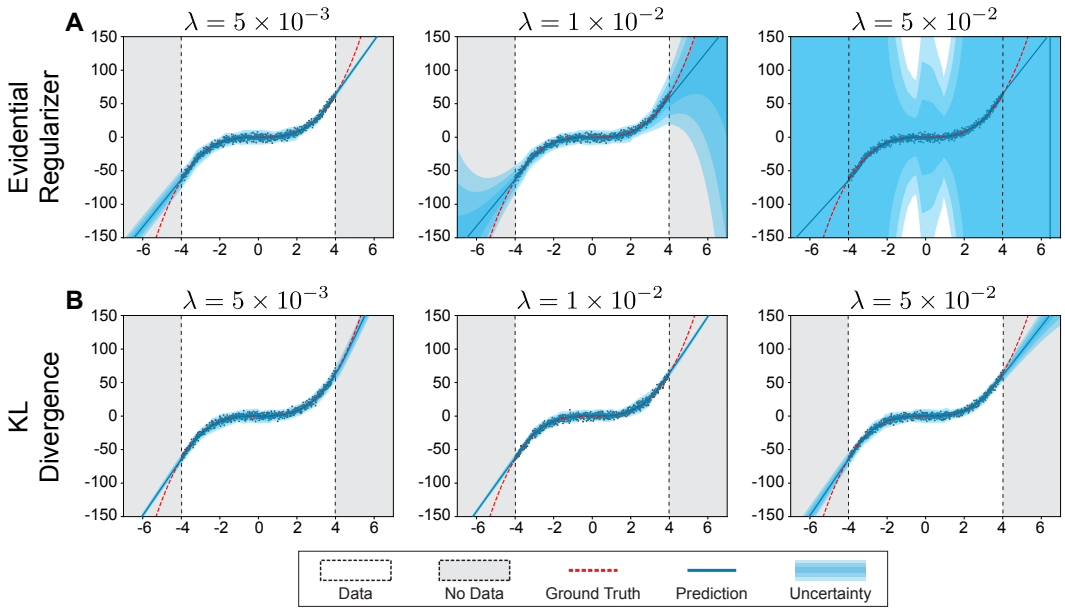

Figure S2: **Impact of regularization strength on epistemic uncertainty estimates.** Epistemic uncertainty estimates on the dataset $y = x^3 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 3)$ for evidential regression models regularized with the evidential regularizer $\mathcal{L}^R$ (A) or with the KL divergence (B) between the inferred `NIG` and another with zero evidence, for varying regularization coefficients λ.

we construct a synthetic toy dataset with high data noise (aleatoric uncertainty) in the center of the in-distribution region. Rather than using the $L1$ error in the regularization term, as in previous experiments, we use regularize the standard score and estimate epistemic and aleatoric uncertainty (Fig. S3). This analysis suggests that the method is capable of disentangling epistemic and aleatoric uncertainties in a region that is in-distribution but has high data noise.

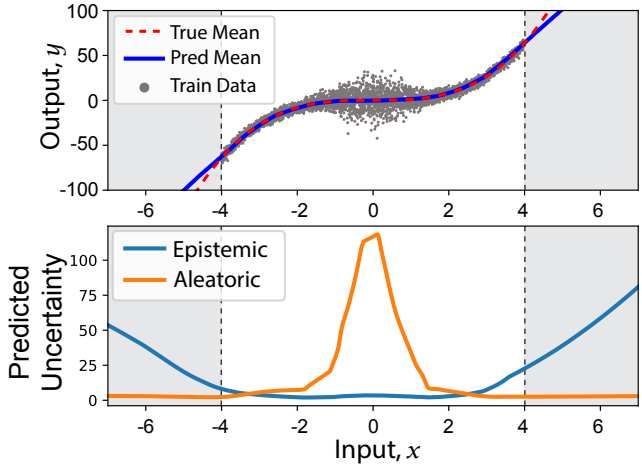

Figure S3: **Disentangled uncertainties.** Epistemic and aleatoric uncertainty estimates on a synthetic dataset based on $y = x^3$, where data noise increases towards the center of the in-distribution region. The evidential regularizer $\mathcal{L}^R$ is calculated based on the standard score.

S2.2 Benchmark regression problems

S2.2.1 Datasets and experimental setup

This subsection describes the setup to create Table 1. We follow an identical experimental setup and training process as presented in [20]. All dataset features are normalized to have zero mean and unit standard deviation. Features with no variance are only normalized to have zero mean. The same normalization process is also performed on the target variables; however, this is undone at

inference time such that predictions are in the original scale of the targets. Datasets are split randomly into training and testing sets a total of 20 times. Each time we retrain the model and compute the desired metrics (RMSE, NLL, and speed). The results presented in the table represent the average and standard error across all 20 runs for every method and dataset. Following the lead of [28], we also directly compare against the other training methods by directly using their reported results since they followed an identical training procedure.

S3 Depth estimation evaluations

S3.1 Experimental details

We evaluate depth estimation on the NYU-Depth-v2 dataset [35]. For every image scan in the dataset we fill in the missing holes in the depth using the Levin Colorization method. The resulting depth map is converted to be proportional to disparity by taking its inverse. This is common in depth learning literature as it ensures that far away objects result in numerically stable neural network outputs (very large depths have close to zero disparity). Objects closer than 1/255 meters to the camera would therefore be clipped due to the `uint8` restriction on image precision. The resulting images are saved and used for supervising the learning algorithm. Training, validation, and test sets were randomly split (80-10-10) with no overlap in scans.

All trained depth models have a U-Net [41] backbone, with five convolutional and pooling blocks down (and then back up). The input and target images had shape $(160, 128)$ with inputs having 3 feature maps (RGB), while targets only had a single feature map (disparity). The dropout variants were trained with spatial dropout [45] over the convolutional blocks ($p = 0.1$). Evidential models additionally had four output target maps, one map corresponding to each evidential parameter $\gamma, \upsilon, \alpha, \beta$, with activations as described in 3.3.

All models were trained with the following hyperparmeters: batch size of 32, Adam optimization with learning rate 5e-5, over 60000 iterations. The best model according to validation set RMSE is saved and used for testing. Evidential models additionally had $\lambda = 0.1$. Each model was trained 3 times from random initialization to produce all presented results.

S3.2 Depth estimation performance metrics

Table S1 summarizes the size and speed of all models. Evidential models contain significantly fewer trainable parameters than ensembles (where the number of parameters scales linearly with the size of the ensemble). Since evidential regression models do not require sampling in order to estimate their uncertainty, their forward-pass inference times are also significantly more efficient. Finally, we demonstrate comparable predictive accuracy (through RMSE and NLL) to the other models.

	N	# Parameters		Inference Speed		RMSE	NLL
		Absolute	Relative	Seconds	Relative		
Evidential (Ours)	-	7,846,776	1.00	0.003	1.00	0.024 ± 0.032	-1.128 ± 0.290
Spatial Dropout	2	7,846,657	1.00	0.028	10.20	0.033 ± 0.037	-0.564 ± 0.231
Spatial Dropout	5	7,846,657	1.00	0.031	11.48	0.031 ± 0.033	-1.227 ± 0.374
Spatial Dropout	10	7,846,657	1.00	0.037	13.69	0.035 ± 0.042	-1.139 ± 0.379
Spatial Dropout	25	7,846,657	1.00	0.065	23.99	0.032 ± 0.035	-1.137 ± 0.327
Spatial Dropout	50	7,846,657	1.00	0.107	39.36	0.032 ± 0.036	-1.110 ± 0.381
Ensembles	2	15,693,314	2.00	0.005	1.94	0.026 ± 0.032	-1.080 ± 3.334
Ensembles	5	39,233,285	5.00	0.010	3.72	0.023 ± 0.027	-1.077 ± 0.298
Ensembles	10	78,466,570	10.00	0.019	6.82	0.025 ± 0.038	-0.980 ± 0.298
Ensembles	25	196,166,425	25.00	0.045	16.45	0.022 ± 0.029	-1.000 ± 0.259
Ensembles	50	392,332,850	50.00	0.112	41.26	0.022 ± 0.031	-0.996 ± 0.275

Table S1: **Depth estimation performance metrics.** Comparison of different uncertainty estimation algorithms and predictive performance on an unseen test set. Dropout and ensembles were sampled N times on parallel threads. The evidential method outperforms all other algorithms in terms of space (#Parameters) and inference speed while maintaining competitive RMSE and NLL.

S3.3 Epistemic uncertainty estimation on depth

Fig. S4 shows individual trial runs for each method on RMSE cutoff plots as summarized in Fig. 4B.

Fig. S5 shows individual trial runs for each method on their respective calibration plots as summarized in Fig. 4C.

Fig. S6 shows individual trial runs for each method on their respective entropy (uncertainty) CDF as a function of the amount of adversarial noise. We present the evidential portion of this figure in Fig. 6C, but also provide baseline results here.

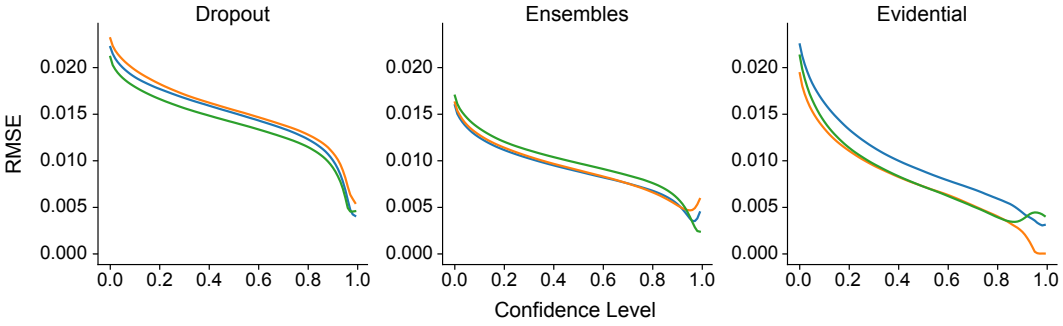

Figure S4: **Relationship between prediction confidence level and observed error for different uncertainty estimation methods.** A strong inverse trend is desired to demonstrate that the uncertainty estimates effectively capture accuracy. Plots show results from depth estimation task.

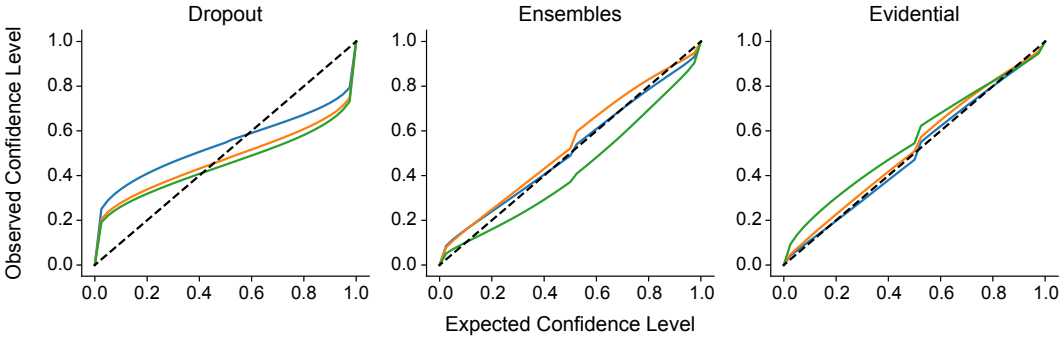

Figure S5: **Uncertainty calibration plots for depth estimation.** Calibration of epistemic uncertainty estimates for dropout, ensembling, and evidential methods, assessed as the relationship between expected and observed predictive confidence levels. Perfect calibration corresponds to the line $y = x$ (black).

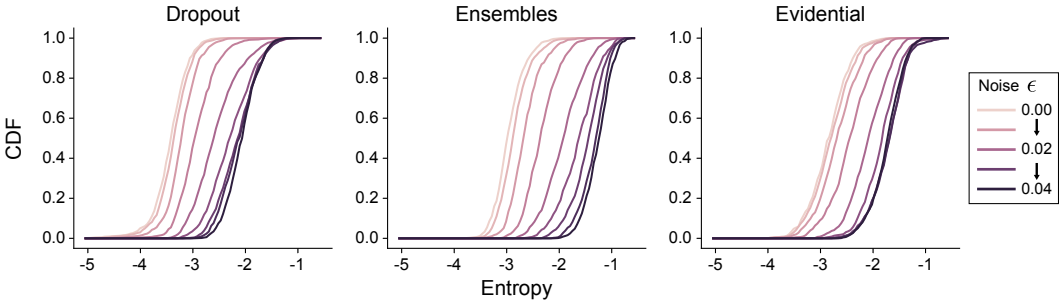

Figure S6: **Effect of adversarial noise on uncertainty estimates**. Cumulative distribution functions (CDF) of entropy (uncertainty) estimated by dropout, ensembling, and evidential regression methods, under the presence of increasing adversarial noise ϵ.

S3.4 Aleatoric uncertainty estimation on depth

Fig. S7 compares the evidential aleatoric uncertainty to those obtained by Gaussian likelihood optimization in several domains with high data uncertainty (mirror reflections and poor illumination). The results between both methods are in strong agreement, identifying mirror reflections and dark regions without visible geometry as sources of high uncertainty. These results are expected since evidential models fit the data to a higher-order Gaussian distribution and therefore it is expected that they can accurately learn aleatoric uncertainty (as is also shown in [42, 18]). While the main text

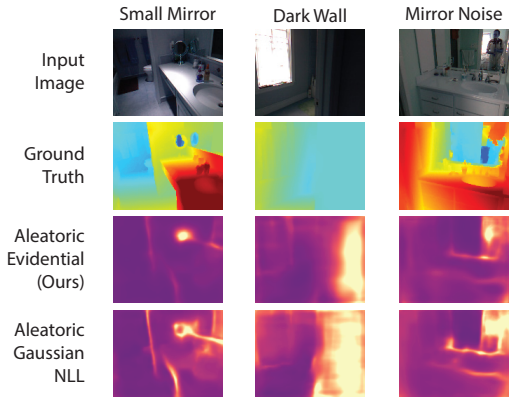

Figure S7: **Aleatoric uncertainty in depth.** Visualizing predicted aleatoric uncertainty in challenging reflection and illumination scenes. Comparison between evidential and [25] show strong semantic agreement.

focuses on the more challenging problem of epistemic uncertainty estimation (especially on OOD data), we provide these sample aleatoric uncertainty examples for here for depth as supplemental material.