Plausible Uncertainties for Human Pose Regression

Lennart Bramlage[1] Michelle Karg[2]
Cristóbal Curio[1]
[1] Cognitive Systems Group, Reutlingen University, Germany
[2] Continental AG
{lennart.bramlage, cristobal.curio}@reutlingen-university.de
michelle.karg@continental.com

Abstract

Human pose estimation (HPE) is integral to scene understanding in numerous safety-critical domains involving human-machine interaction, such as autonomous driving or semi-automated work environments. Avoiding costly mistakes is synonymous with anticipating failure in model predictions, which necessitates meta-judgments on the accuracy of the applied models. Here, we propose a straightforward human pose regression framework to examine the behavior of two established methods for simultaneous aleatoric and epistemic uncertainty estimation: maximum a-posteriori (MAP) estimation with Monte-Carlo variational inference and deep evidential regression (DER). First, we evaluate both approaches on the quality of their predicted variances and whether these truly capture the expected model error. The initial assessment indicates that both methods exhibit the overconfidence issue common in deep probabilistic models. This observation motivates our implementation of an additional recalibration step to extract reliable confidence intervals. We then take a closer look at deep evidential regression, which, to our knowledge, is applied comprehensively for the first time to the HPE problem. Experimental results indicate that DER behaves as expected in challenging and adverse conditions commonly occurring in HPE and that the predicted uncertainties match their purported aleatoric and epistemic sources. Notably, DER achieves smooth uncertainty estimates without the need for a costly sampling step, making it an attractive candidate for uncertainty estimation on resource-limited platforms.

1. Introduction

As more and more deep learning-based machine vision systems are increasingly utilized in safety-critical real-world applications, the need for principled safeguards becomes impossible to ignore. Human pose estimation (HPE) is an application that allows autonomous systems to anticipate peoples' intentions and movements and avoid consequential mistakes. Every model is flawed, and decision-making under uncertainty presupposes robust uncertainty estimation. A vast catalogue of methods attempts to meet this requirement and augment point-wise model predictions with measures of uncertainty and confidence. However, while drawing on rigorous foundations, strong theoretical assumptions generally need to be relaxed to meet the realities of deployment. For example, "irreducible" aleatoric uncertainty, in many real-world applications, reduces to conditional variance. Despite such concessions, established uncertainty estimation and disentanglement methods may harbor exceptional potential, provided they behave as expected in any given domain. Here, we investigate whether two es-

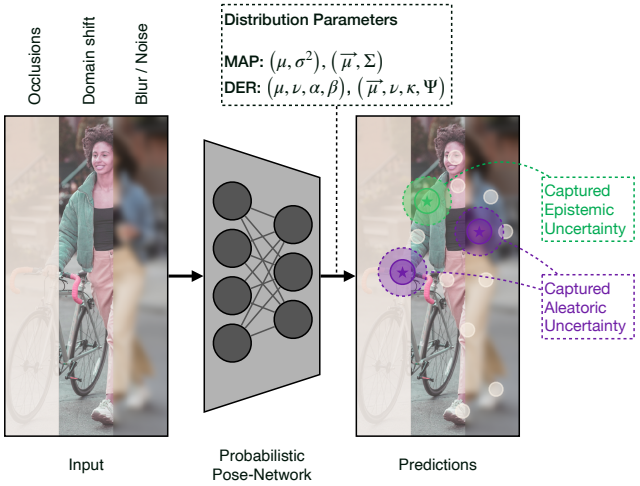

Figure 1. We employ two methods of uncertainty quantification to identify the aleatoric and epistemic components of the expected error in human pose estimation, anticipating large aleatoric uncertainty in occluded and blurred joints and increased epistemic uncertainty in novel domains.

tablished methods of uncertainty quantification remain true to their definitions in the challenging HPE domain, specifically:

1. We present two methods for simultaneous aleatoric and epistemic uncertainty quantification, maximum a-posteriori estimation (MAP) and deep evidential regression (DER).
2. We motivate a recalibration step for robust estimation of confidence intervals in order to derive plausible and interpretable measures of confidence.
3. To the best of our knowledge, this is the first full-fledged application of evidential deep learning to the HPE domain. For this reason, we additionally conduct thorough studies on the interpretability of the results including label noise injection and occlusion trials.

1.1. Why estimate two types of uncertainty?

Imagine an unfair coin toss: knowing whether the coin is weighted and to what degree will significantly reduce uncertainty about which side might come up more frequently; however, even a weighted coin might come up on the other side with some frequency (aleatoric). Reducing uncertainty about our model of the coin's characteristics (epistemic) improves the quality of our predictions, but it will not make them perfect. An ability to gauge at what point we have learned everything there is to learn about a stochastic system is indispensable for efficient learning and effective risk assessment. Formally, we distinguish *aleatoric* uncertainty, stemming from, e.g., sensor- or measurement noise or the inherent randomness in the data generating process, and *epistemic* uncertainty, also called model-uncertainty, the uncertainty in the model specification or parameters [10, 41]. In vision, examples of aleatoric uncertainty include motion blur, low contrast, or compression artifacts. Another possible source of aleatoric uncertainty in supervised learning may be noise in the labeling process. Common applications of aleatoric noise models are in outlier- or out-of-distribution detection [44], as well as sample rejection [43], and the assessment of expected value and risk [8]. Epistemic uncertainty, on the other hand, can stem from a shift into a previously unseen domain and generally diminishes as more training data becomes available to the model, assuming the optimization procedure is well-defined. In this way, accurate epistemic uncertainty estimates enable various downstream applications, such as active sampling optimization methods [34] and active learning [6]. Even an empirical reevaluation of model choice is possible.

2. Related Work

2.1. Uncertainty Quantification in Deep Learning

Epistemic uncertainty in neural networks can be assessed by placing a Gaussian prior over the model parameters $\mathbf{W} \sim \mathcal{N}(0, I)$ giving rise to a Bayesian neural network (BNN) [29, 7]. Modern large-scale neural networks tend to struggle with high variance over high bias which is why uncertainty in parameters dominates the research, more so than uncertainty about the model specification. In practice, it is impossible to evaluate the posterior over the parameters $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$ since the marginal $p(\mathbf{Y}|\mathbf{X})$ requires integration over all possible sets of model parameters. Various approximate methods exist [11, 4, 19], such as Dropout Variational Inference [10, 9] which models a sample of the weight space by resampling a trained model with intermittent dropout layers during inference. Here, the variance over the aggregated point estimates serves as a realization of epistemic uncertainty. Ensemble methods similarly constitute subspaces of possible models, and treat model disagreement as a measure of epistemic uncertainty [35, 46, 26]. Moreover, in recent years more approaches attempting to estimate epistemic uncertainty directly have emerged [17].

Aleatoric uncertainty is by definition sample-dependent and generally estimated as an auxiliary and predominantly unsupervised task in neural network training. Where epistemic uncertainty arises naturally in approaches like dropout variational inference, aleatoric uncertainty needs to factor into the loss function directly. One widely applied approach is fitting the parameters of a Gaussian instead of a simple point estimate, where the variance, conditioned on the input sample, is considered to be a measure of aleatoric uncertainty. In the general absence of uncertainty-annotated data, prediction errors are weighted by this predicted variance in order to attenuate the loss in high uncertainty regions of the input space [18, 15].

Many alternative approaches exist, such as the direct construction of confidence and prediction intervals [36, 23] or conformal prediction [39], a distribution-free method to generate sets of neural network outputs that are guaranteed to contain the true value with a given probability. Lastly, evidential deep learning methods can infer both types of uncertainty in a single forward pass by placing a prior distribution over the likelihood function [38, 3, 1]. The variance of the prior encapsulates a notion of epistemic uncertainty. Sampling this prior leads to a single realization of the likelihood function that, in turn, can be sampled to generate a single point estimate with an associated variance, i.e., a measure of aleatoric uncertainty.

2.2. Uncertainty Quantification in HPE

A common benchmark paradigm in HPE, heatmap regression, inserts a degree of uncertainty by using heatmaps with Gaussian noise around the ground-truth joint-data as training targets [42]. This method is generally more robust than a direct regression of joint-coordinates. An obvious shortcoming is that the noise is inserted manually during the labeling process and subject to a design choice. Fur-

ther, it tends to be homoscedastic across the data, which makes output heatmaps unfit to model either of the discussed types of uncertainty directly. Additionally, predicting entire heatmaps instead of single coordinates adds significant time- and space-complexity to training and inference. Due to its application in safety critical real-world problems, HPE has thus witnessed renewed interest in the study of efficient and robust uncertainty estimation.

A variety of methods aim to augment the heatmap paradigm with provisions for uncertainty estimation. [22] define uncertainty as a disagreement between model-based and model-free pose-estimation heads in order to improve out-of-distribution performance of their model. [28] instead propose two heads to output heatmaps and scale maps, where scale maps are used to scale each sample's ground-truth heatmap during training, inducing a weight-adaptive loss function. However, these approaches comprise not only costly heatmap regression, but also a multitude of specialized sequences and nodes for inference. [24] propose a straightforward regression-based method that estimates complex distributions of deviations from the ground truth, conditioned on input images, using normalizing flows. Intuitively, these distributions could be described as a measure of aleatoric uncertainty but while performance results are promising, the uncertainty perspective is not investigated further. [12] make use of the popular MAP method introduced by [18] to fit a multivariate Gaussian distribution over joint locations. They demonstrate improved performance over the traditional heatmap-based approach while focusing entirely on the aleatoric uncertainty aspect. [13] likewise consider only the aleatoric uncertainty in their adaptation of the MAP method for HPE. They generate a single variance estimate for each joint and go on to employ these estimates in a graph neural network to refine pose predictions.

3. Method

Fig. 2 provides an overview of the proposed approach. Keypoint localization for body pose estimation is extended by uncertainty quantification. Two different methods are benchmarked for 2d and 3d keypoint estimation for datasets differing in scene, body poses, real/simulated/lab setting, resolution, and ground truth accuracy.

3.1. Intuition

Established uncertainty quantification methods model the expected error as a combination of aleatoric and epistemic uncertainty, i.e., as a sum of the inherent randomness in the data-generating process and the model's lack of capacity to capture the target value. While the epistemic element translates reasonably well to the unstructured HPE domain, the aleatoric uncertainty may appear less obvious. In low-dimensional regression problems, capturing the aleatoric uncertainty reduces to the maximum likelihood estimation formulation of the variance conditioned on some value or interval of the input space. In HPE, the aleatoric uncertainty could be defined as the predicted *joint-ness* of an image region under the estimated aleatoric variance. An increase in size of this region corresponds directly to higher expected error in the joint location estimate.

3.2. Uncertainty Quantification

We compare two methods of predicting aleatoric and epistemic uncertainties simultaneously. The first is well-established and combines per-sample MAP with variational inference through Monte-Carlo dropout (MCD) to predict both types of uncertainty [18]. The second method, called deep evidential regression (DER), wraps both types of uncertainty into a single loss function, and therefore does not require a costly sampling step [1]. We will illustrate both methods in detail.

Maximum A-Posteriori Inference: A widely established method of estimating heteroscedastic uncertainty is via direct estimation of the parameters of a Gaussian distribution, conditioned on the input data, such that $\mathbf{f}_\theta : \mathcal{X} \mapsto \{\hat{\mu}, \hat{\sigma}^2\}$. The corresponding minimization objective for an $L2$-loss is the negative log-likelihood of the Gaussian:

$$\begin{aligned}\mathcal{L}_{nll}(\theta) &= -\log p(y|\hat{\mu}, \hat{\sigma}^2) \\ &= \frac{1}{K}\frac{1}{D}\sum_k \sum_d \frac{\|y_{kd} - \hat{\mu}_{kd}\|^2}{2\hat{\sigma}_{kd}^2} + \frac{1}{2}\log \hat{\sigma}_{kd}^2 \end{aligned} \quad (1)$$

where D is the number of output dimensions and K the number of joints, in our case $D = 2$ or $D = 3$ for 2d- and 3d-pose estimation respectively, while K depends on the available ground-truth data for each sample. Deriving the negative log-likelihood of the Gaussian presents a well-balanced convex optimization problem. The squared error is attenuated by the estimated variance parameter, however, the added log-variance term prevents this parameter from growing infinitely. Effectively, the model can mitigate large errors by correspondingly large variance estimates. Hence, the overall loss can be further reduced on small errors by estimating small variances. As suggested by [18], a more stable variant of this loss function assumes the estimation of the log-variance. We further choose an $L1$-loss for our approach, following prior work on human pose regression and generally better performance [42, 15]. We adapt the above loss function accordingly and fit a Laplace distribution as follows:

$$\mathcal{L}_{nll}(\theta) = \frac{1}{K}\frac{1}{D}\sum_k \sum_d \exp(-b_{kd})|y_{kd} - \hat{\mu}_{kd}| + b_{kd} \quad (2)$$

where $2b_{kd}^2 := \log \hat{\sigma}_{kd}^2$. MCD can be considered variational inference to estimate a simpler distribution over model parameters $q^*(\mathbf{W})$ that minimizes the KL-divergence to the intractable model posterior $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$. The variance of

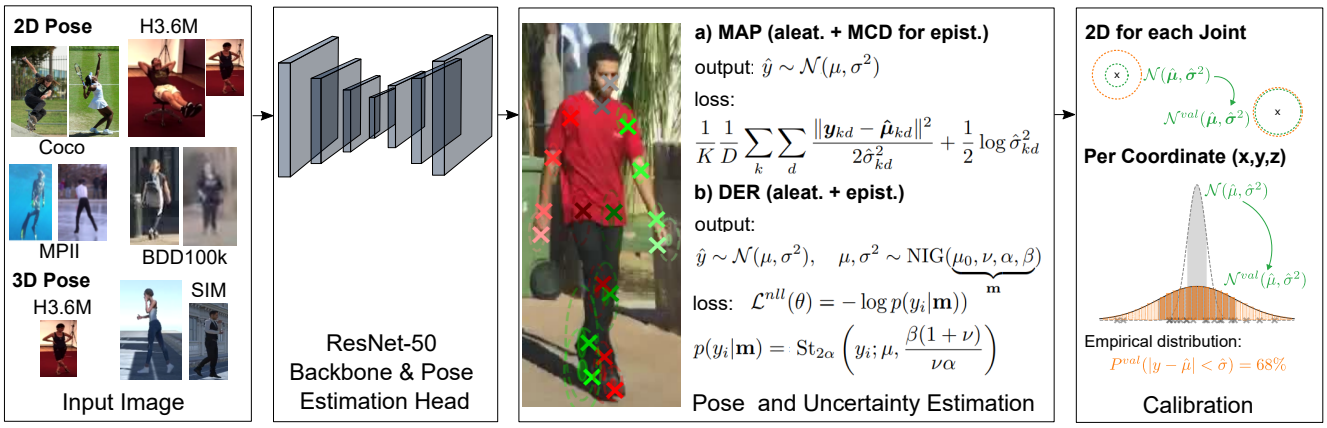

Figure 2. We compare the MAP + MCD and the DER approach for estimating both the aleatoric and the epistemic uncertainty of each joint location. The estimated uncertainties are calibrated based on the empirical distribution of the residuals for the validation dataset. The approach is evaluated for four 2d body pose datasets and two 3d body pose datasets.

this approximation yields the epistemic uncertainty estimate. We resample the final three network layers of the respective 2d- and 3d-pose heads with intermittent MCD 50 times and derive the epistemic uncertainty as $\hat{\sigma}^2_{MC} = \text{Var}[\hat{\mu}_{kd}]$ with $\hat{\sigma}^2_{MC} \in \mathbb{R}^{+, D\times K}$.

Deep Evidential Regression: DER implements simultaneous estimation of aleatoric and epistemic uncertainty by fitting the parameters of the conjugate prior of a Gaussian distribution with unknown mean μ and variance σ^2, i.e., the Normal-Inverse Gamma (NIG) distribution, such that $\mathbf{f}_\theta : \mathcal{X} \mapsto \{\hat{\mu}, \hat{\nu}, \hat{\alpha}, \hat{\beta}\}$. Drawing samples from this prior yields individual instances of the likelihood function, which can then be sampled to receive the network prediction.

$$\hat{y} \sim \mathcal{N}(\mu, \sigma^2), \quad \mu, \sigma^2 \sim \text{NIG}(\underbrace{\hat{\mu}_0, \hat{\nu}, \hat{\alpha}, \hat{\beta}}_{\mathbf{m}}) \tag{3}$$

By marginalizing over the likelihood function parameters, we obtain the predictive posterior, i.e., the posterior probability of the target value, given the parameters of the prior. This is a Student's t-distribution in the form of a three-parameter location-scale distribution with 2α degrees of freedom, location μ, and scale $\frac{\beta(1+\nu)}{\nu\alpha}$ (see Sec. A for details). Hence, the loss function is:

$$\begin{aligned}\mathcal{L}_{nll}(\theta) &= -\log p(y|\mathbf{m}) \\ &= -\log \int_{\sigma^2=0}^{\infty}\int_{\mu=-\infty}^{\infty} p(y|\mu,\sigma^2)p(\mu,\sigma^2|\mathbf{m})\mathrm{d}\mu\,\mathrm{d}\sigma^2 \\ &= \frac{1}{K}\frac{1}{D}\sum_K\sum_D -\log \text{t}_{2\alpha}\left(y_{kd}; \hat{\mu}_{kd}, \frac{\hat{\beta}_{kd}(1+\hat{\nu}_{kd})}{\hat{\nu}_{kd}\hat{\alpha}_{kd}}\right)\end{aligned} \tag{4}$$

In addition to the negative log-likelihood, [1] suggest to add a regularization term punishing large estimates of virtual evidence when the error is large, resulting in the following objective function:

$$\mathcal{L}(\theta) = \mathcal{L}_{nll}(\theta) + \lambda|y_{kd} - \hat{\mu}_{kd}|(2\hat{\alpha}_{kd} + \hat{\nu}_{kd}) \tag{5}$$

The desired uncertainties can subsequently be derived from the predicted NIG parameters as follows:

$$\underbrace{\mathbb{E}[\sigma^2] = \frac{\beta}{(\alpha-1)}}_{\text{aleatoric}}, \quad \underbrace{\text{Var}[\mu] = \frac{\beta}{\nu(\alpha-1)}}_{\text{epistemic}} \tag{6}$$

3.3. The Multivariate Case

To consider interactions between the x and y coordinate for 2D pose estimation and between x, y, and z coordinates for 3D pose estimation, we additionally investigate multivariate variants of the DER and MAP approach. The joints remain independent. For clarity, we omit the k subscript indicating the joint.

Maximum A-Posteriori Inference: We fit the negative log-likelihood of the multivariate Gaussian as the objective function for each joint k:

$$\begin{aligned}\boldsymbol{z} &= (\boldsymbol{y}_k - \hat{\boldsymbol{\mu}}), \\ -\log p(\boldsymbol{y}_k|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= \frac{1}{2}\left[\boldsymbol{z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{z} + \log\det(\hat{\boldsymbol{\Sigma}})\right]\end{aligned} \tag{7}$$

This objective requires the prediction of the invertible covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{+n\times n}$ and may therefore lead to numerical instabilities during optimization. We can instead estimate the inverse, i.e., the precision matrix $\boldsymbol{\Psi}$ directly, leading to the negative log-likelihood with respect to $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1}$:

$$\mathcal{L}_{nll}(\theta) = \frac{1}{2}\mathbb{E}\left[\boldsymbol{z}^\top\hat{\boldsymbol{\Psi}}\boldsymbol{z} - \log\det(\hat{\boldsymbol{\Psi}})\right] \tag{8}$$

$\boldsymbol{\Psi}$ is positive definite with elements $\boldsymbol{\Psi}_{ii} > 0$, which is guaranteed by applying the Cholesky Decomposition and estimating the lower triangular matrix $\hat{\boldsymbol{L}}$ instead, where

$$\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^\top, \text{ with } \hat{\boldsymbol{L}}_{ii} > 0. \tag{9}$$

Analogously to the univariate case, we estimate the epistemic component using MCD and the sample covariance of 50 samples.

Deep Evidential Regression: Similarly, for the multivariate DER, we derive the negative log-likelihood of the multivariate Student's t-distribution, parameterized in terms of the parameters $\boldsymbol{m} = \{\boldsymbol{\mu}_0, \nu_0, \kappa, \boldsymbol{\Psi}\}$ of the conjugate prior of a multivariate Gaussian, the Normal-Inverse Wishart distribution [33] (see Sec. B for derivation).

$$\mathcal{L}_{nll}(\theta) = -\log \mathrm{t}_{\hat{\nu}_0 - n + 1}\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0, \frac{\hat{\kappa}+1}{\hat{\kappa}(\hat{\nu}_0 - n + 1)}\hat{\boldsymbol{\Psi}}\right) \tag{10}$$

Where n is the number of covariates and $\hat{\boldsymbol{\Psi}}$ is the estimated sum of squared errors, allowing us to infer aleatoric uncertainty, according to the definition of the mean of the Inverse-Wishart distribution, via $\mathbb{E}[\boldsymbol{\Sigma}] = \boldsymbol{\Psi}/(\nu_0 - n - 1)$ [30] and epistemic uncertainty as $\mathrm{Var}[\boldsymbol{\mu}] = \mathbb{E}[\boldsymbol{\Sigma}]/\nu_0$ [32]. [32] observe, that the κ parameter is expected to converge on 1, due to the definition of the evidential loss function. We choose not to adopt their suggested solution of linearly coupling the two scalar parameters $r\hat{\kappa} = \hat{\nu}$ because it leads to uncalibrated uncertainty estimates.

3.4. Accurate Uncertainties with Calibrated Regression

For regression tasks, a forecaster H is calibrated if

$$\frac{\sum_{n=1}^{N} 1\{y_n \leq F_n^{-1}(p)\}}{N} \rightarrow p, \forall p \in [0,1] \tag{11}$$

as $N \rightarrow \infty$. $F_n^{-1} : [0,1] \mapsto \mathcal{Y}$ is the learned quantile function conditioned on sample n [21]. That is, the fraction of values y_n inside the pth quantile of the quantile function derived from $\{\hat{\mu}_n, \hat{\sigma}_n^2\}$ should be equal to p as the size of the dataset approaches infinity. This is usually not the case for deep probabilistic models, hence, the predicted confidence intervals require recalibration, a rescaling, to match the empirical cumulative density function (CDF). For recalibration, we create a recalibration dataset $\mathcal{D}$ as follows:

$$\mathcal{D} = \left\{\left(F_n(y_n), \hat{P}(F_n(y_n))\right)\right\}_{n=1}^{N},$$
$$\text{where} \tag{12}$$
$$\hat{P}(p) = \frac{|\{y_n | F_n(y_n) \leq p, n, ..., N\}|}{N}$$

Here, $F_n(y_n) : \mathcal{R} \mapsto [0,1]$ is the predicted CDF, conditioned on input sample x_n, evaluated at the ground truth value y_n. $\hat{P}(p)$ denotes the empirical CDF, the fraction of the data for which y_n lies below the pth quantile of F_n, our predicted CDF. For a sharp forecaster, and as $N \rightarrow \infty$, $\hat{P}(p) \rightarrow p, \forall p \in [0,1]$. We subsequently use the recalibration dataset to fit an auxiliary isotonic regression model $R : [0,1] \mapsto [0,1]$ and output calibrated quantiles [21, 20]. We can now generate robust confidence intervals for unseen values by scaling the predicted standard deviations for any desired quantile:

$$\hat{F}_n^{-1}(p) = F_1^{-1}(R(p)) * \hat{\sigma}_n \tag{13}$$

where F_1^{-1} is the quantile function of the standard normal distribution. Recalibration occurs on a per joint and per coordinate basis and conversely induces an adapted CDF and PDF as shown in Fig. 3.

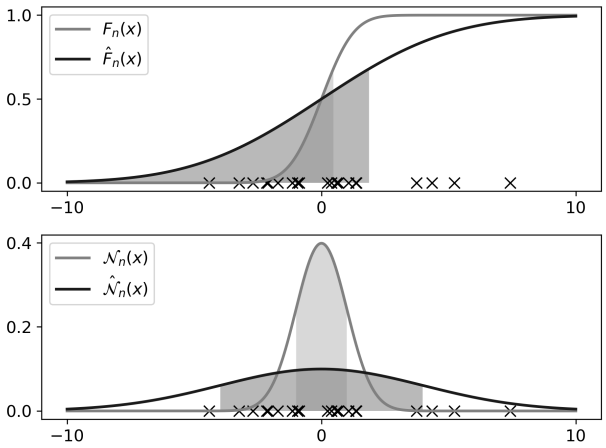

Figure 3. Recalibrated confidence scores effectively rescale the predicted distribution to better match the data. Here, a $\hat{\cdot}$ denotes the resulting CDF and PDF derived from recalibrated confidence intervals.

4. Experiments

4.1. A Minimal Prediction Head

Each of our models consists of a ResNet-50 [14] backbone coupled with separate feature heads for 2d- and 3d-pose estimation. The feature heads comprise one convolutional layer, as well as two fully connected layers and a final layer to generate the respective distribution parameters. Unlike prior works, such as [24], the final, fully connected layer comprises only 4 (MAP), 8 (DER), $2+3$ ((mv) MAP), $4+3$ ((mv) DER) neurons in the two-dimensional case. All fully connected layers are applied batchwise across the channels of the initial convolutional layer, which reflect the maximum number of joints $K = 26$ across all datasets. Combining this simplification with a smaller number of 1024 units in the penultimate layer, we significantly reduce the number of weights in the final regression layer.

4.2. General Training

We train each model for 80 epochs, using MS-COCO [25], H36M [16], MPII [2], and SIM during the 10 fi-

	DER		MAP		
pck02 ↑	uni	multi	uni	multi	IHPR
MS-COCO [25]	89.2	**92.9**	88.9	88.5	87.2
H36M [16]	92.3	92.2	**92.5**	92.3	90.3
MPII [2]	86.2	**88.5**	85.5	85.3	81.6
SIM	99.2	**99.8**	99.5	99.1	95.0
*BDD100k [45]	65.2	**66.5**	65.7	64.7	55.2
mpjpe ↓					
H36M [16]	73.7	**68.9**	74.4	73.9	109.8
SIM	14.5	28.2	15.7	**14.3**	23.0

Table 1. 2d- and 3d-performance metrics for *uni-* and *multi*variate versions of the proposed methods. All scores are averaged over coordinates and joints, supported by the dataset in question. The BDD100k dataset is unseen during training.

nal epochs. For the MS-COCO AP evaluation, we train the model on MS-COCO exclusively for a final step. We employ typical protocols for each of the datasets, such as the MS-COCO 2017 train/test/val split, training on subjects (S1, S5, S6, S7, S8) and testing on (S9, S11) for H36M, and training on 25k out of 40k human pose samples from MPII. The SIM dataset was created internally, using a simulated street scene, 11 camera views, and 3d-scanned subjects (see Sec. C for details) [5]. SIM serves as an excellent testbed for aleatoric uncertainty estimation, since its labels are consistent across all samples. Lastly, we evaluate all models on the BDD100k dataset [45], which remains unseen during training, to gauge out-of-distribution performance and calibration. The input size for all models is 256×192 and we train with a batch size of 48. All datasets are subject to the same augmentation pipeline of random flips with $p = 0.5$, normally distributed random rotations $\in [-30°, 30°]$, and random rescaling by a factor $\in [0.5, 1.5]$. We use the *AdamW* optimizer [27] with *1cycle* [40] learning rate policy at a maximum learning rate of $1e^{-3}$. As suggested by prior work [10], we employ weight decay to stabilize uncertainty estimates with $\lambda = 1e^{-2}$ and find further that weight decay as low as $\lambda = 1e^{-4}$ leads to calibrated uncertainties in the case of DER. Additionally, we provide a baseline comparison using the heatmap-based Integral Human Pose Regression method (IHPR) [42]. As mentioned in the introductory sections, this model class handles various sources of uncertainty reasonably well without generating plausible uncertainty estimates and, hence, we only enter it in the general task performance comparison.

4.3. Task Performance

We use **pck02** (percentage of predicted points within 0.2 times the torso diameter distance to the correct keypoint) to assess 2d-pose quality and **mpjpe** (mean per-joint position error, Euclidean distance between true and predicted joint) to evaluate 3d-pose quality. In terms of general performance, the models are virtually identical both in the 2d- and 3d-case, see Tab. 1. The baseline performs slightly worse than average which is to be expected given that we employ IHPR, whose heatmaps cannot capture and thus do not account for uncertainties in the input. Importantly, since the DER models did not require resampling of multiple layers with intermittent dropout, training and inference were up to three times faster when compared with MAP on our system (see details in Sec. D). Tab. 2 shows AP performance results on the MS-COCO validation set. While our model does not manage to outperform competing approaches that incorporate uncertainty estimation, it manages to reach comparable performances, even with the above-stated minimal prediction head. It also manages to do this while simultaneously estimating two types of uncertainty, which other approaches neglect to implement.

Method	Backbone	AP	AP_{50}	AP_{75}
RLE [24]	ResNet-50	**70.5**	88.5	**77.4**
SWAHR [28]	HRNet-w32	67.9	**88.9**	74.5
(mv) DER	ResNet-50	67.8	88.5	72.2

Table 2. Comparison with prior works incorporating an uncertainty estimation approach on the MS-COCO validation set.

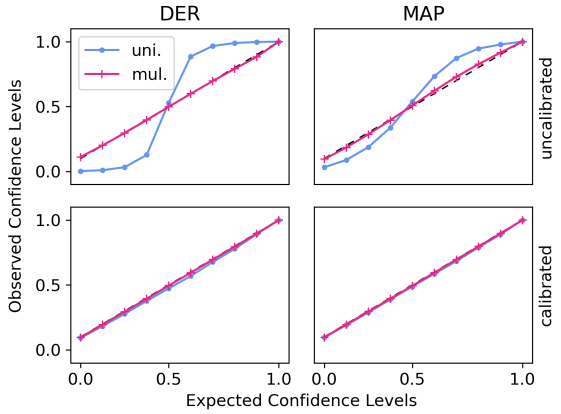

Figure 4. Calibration plots before and after recalibration, based on respective uncertainty estimates, averaged across all datasets and joints. Note that post-recalibration plots are difficult to tell apart because of their proximity.

4.4. Calibration

In order to determine whether the predicted uncertainties hold any relevant information about the expected error, we assess the calibration of each model. Perfect calibration assumes that exactly fraction p of the data fall into the pth confidence interval for all intervals p, see the methods section for the full definition. We can examine this criterion by computing the expected calibration error (ECE) for each

	aleatoric $\mathbb{E}[\sigma^2]$				epistemic $\mathrm{Var}[\mu]$			
ECE ↓	DER	(mv) DER	MAP	(mv) MAP	DER	(mv) DER	MAP	(mv) MAP
MS-COCO [25]	.204 / .025	.065 / .007	.195 / .033	*.014* / **.004**	.054 / **.006**	.041 / .008	*.031* / .008	.034 / .008
H36M [16]	.158 / **.016**	.090 / .021	.198 / .037	*.087* / .027	.091 / .035	.034 / .012	*.033* / .037	.036 / **.006**
MPII [2]	.207 / .023	*.005* / **.003**	.194 / .033	.008 / .004	.053 / .006	.038 / **.004**	*.036* / .007	.040 / .009
SIM	.166 / .031	*.089* / .032	.206 / .039	.126 / **.030**	.117 / .024	*.097* / **.021**	.135 / .033	.127 / .022
*BDD100k [45]	.153 / .016	.047 / .011	.183 / .020	*.034* / **.008**	*.060* / **.018**	.065 / .023	.115 / .044	.113 / .044

Table 3. Aleatoric and epistemic expected calibration error, before $(\cdot/)$ and after $(/\cdot)$ recalibration using isotonic regression. As expected, recalibration significantly improves calibration error across all models and datasets.

$\rho \uparrow$ Method	al.	ep.
DER	**0.87**	0.62
(mv) DER	0.76	0.66
MAP	0.67	0.55
(mv) MAP	0.76	**0.92**
RLE ($Q(\bar{x}) \sim$ Laplace) [24]	0.55	-

Table 4. Pearson correlation coefficient between estimated uncertainty and true residual for the MS-COCO validation set.

model, dataset, joint, and type of uncertainty.

Tab. 3 shows the aleatoric ECE of all models in all datasets. Before recalibration using isotonic regression, both univariate models are approximately on par. The multivariate models outperform both, before and after recalibration, suggesting that a covariate perspective holds value when considering ostensibly independent joint coordinates. The observation becomes even clearer when considering the calibration plots in Fig. 4, which shows that the multivariate approaches are almost perfectly calibrated (i.e., close to the dashed diagonal) out of the box. The same is true for the epistemic dimension, where the multivariate models outperform the univariate ones as well (see Tab. 3).

4.5. Variance-Residual Correlation

Calibration alone is not a sufficient criterion to judge the quality of a probabilistic forecaster. Since predicted uncertainty can be understood to equal the expected error conditioned on an input, we investigate the correlation between uncertainties and true residuals. We expect this relationship to be monotonic, i.e., as the true error grows, so should the predicted uncertainty. For the sake of comparison to prior works, we provide the Pearson correlation coefficient. However, the Spearman's rank correlation coefficient would constitute a stronger criterion, thus we include a per joint and model table in the supplementary materials (see Sec. G), averaged over joint coordinates and datasets, where -1 or 1 indicates perfect correlation. Tab. 4 shows the computed correlation scores for the MS-COCO dataset. All four models display moderate to strong degrees of correlation between predicted aleatoric and epistemic uncertainties and true residuals, outclassing prior work. We find that the top performer in the aleatoric component is the univariate DER approach, suggesting that the estimated coordinates are largely uncorrelated. By contrast, multivariate MAP leads in the epistemic component, weakening this assumption. Note that we compute correlations for uncertainties and the full residual and have no way of knowing how much of the error can be explained by epistemically reducible factors. In other words, aleatoric factors could be captured by estimation methods for epistemic uncertainty. For this reason, we subject the less-explored DER approach to further analysis and attempt to gauge the degree of disentanglement in the following section.

5. Analysis of DER

To investigate the plausibility of DER's uncertainty estimates further, we conduct a series of experiments geared towards increasing either aleatoric or epistemic estimates exclusively. These experiments reflect common sources of uncertainty in HPE and, if captured correctly, are invaluable to downstream decision-making or pre-deployment design decisions. Unlike prior work, we choose not to employ common adversarial methods to provoke uncertainty estimates, but instead devise naturally occurring challenges in HPE. Due to its stability during training, we limit these experiments to the univariate DER model.

5.1. Label Noise Injection

Label noise, stemming from individual biases in manual labeling, is by all intents and purposes irreducible beyond a certain point and thus a prime candidate to assess the validity of aleatoric estimates. With access to a homogeneously labeled simulated dataset, we can inject artificial labeling noise to, ostensibly, induce higher aleatoric uncertainty estimates in a trained model.

We train DER from scratch, for 15 epochs, on six iterations of our SIM dataset with added independent Gaussian noise, with $\sigma \in [0.0, 0.5]$ in the normalized joint coordinate domain $Y \in [0, 1]$. Each model reaches validation scores of 99% **pck02**, indicating that the added noise has negligible

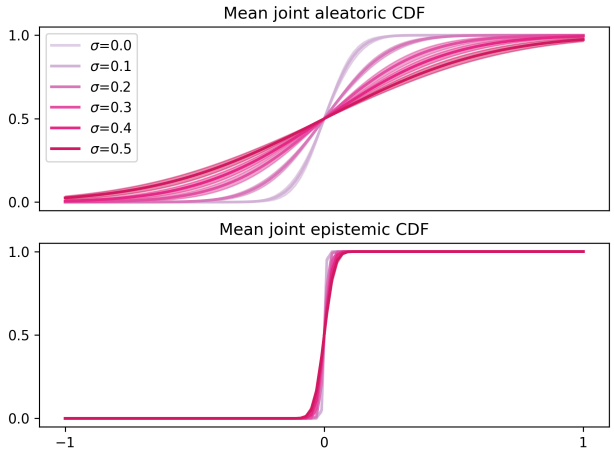

Figure 5. Extracted CDFs averaged across joints, based on aleatoric and epistemic uncertainty estimates after training with varying levels of added Gaussian label noise. Both types of uncertainty increase with added label noise, however, the effect is much more pronounced in the aleatoric component.

effect on overall performance.

We compare average CDFs per joint, drawn from predicted aleatoric and epistemic uncertainty estimates on the validation set, see Fig. 5. The model is successful in capturing a relative increase in aleatoric variance with increased label noise with a negligible increase in epistemic uncertainty. A small amount of aleatoric variance in the zero-noise case can be explained by digital artifacts resulting from upscaling in cases where the bounding box is smaller than the required network input. DER appears to exhibit a common problem in probabilistic deep learning, namely the tendency towards overconfidence since the average predicted variances are generally lower than the added label noise. Curiously, this does not seem to be the case in simpler, one-dimensional regression problems (see Sec. E), further motivating the application of a recalibration step in more complex domains.

5.2. Occlusion

Occlusions can reasonably be considered sources for both types of uncertainty in HPE. Assume an occlusion of the wrist joint with all other joints visible. It would be trivial for an experienced estimator to pinpoint the wrist location accurately. However, if the entire arm were occluded up to the shoulder, the irreducible amount of uncertainty would increase significantly. We place artifical occluders (white, gray, and Gaussian blur) on the wrist joints across all datasets, with occluder sizes equaling a fraction $\phi \in [0.1, 0.5]$ of the ground-truth bounding box diagonal, see Fig. 6. Fig. 7 shows captured uncertainties of a single occluded joint. As expected, increasing occluder size corresponds to increasing aleatoric uncertainty estimates for the

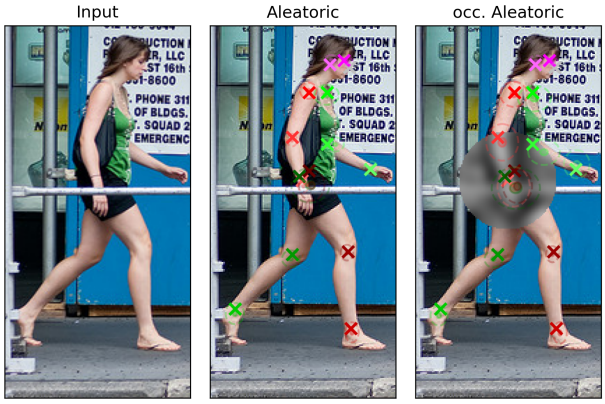

Figure 6. Artificial occlusions increase localized aleatoric uncertainty without increasing epistemic uncertainty estimates.

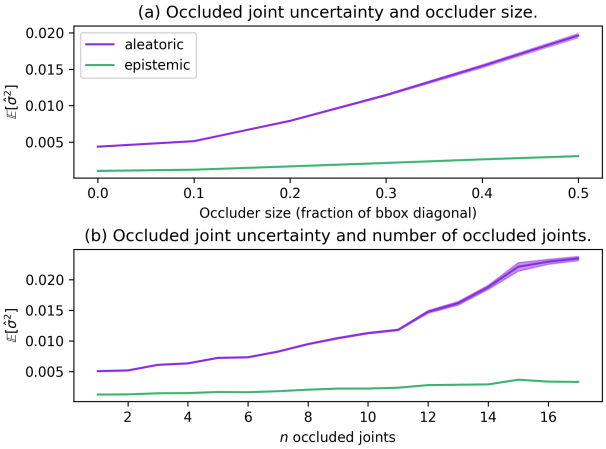

Figure 7. (a) An average increase in predicted uncertainties is positively correlated with the occluder size. (b) The number of collaterally occluded joints exhibits a somewhat weaker correlation.

occluded joint, see Fig. 7 a. We also observe a less significant increase in estimated epistemic uncertainties. Further analysis, however, indicates that epistemic uncertainty displays the same relative increase as aleatoric uncertainty. This may be a consequence of DER's dependence on the single ν parameter to disentangle both types of uncertainty, see Eq. (6), a shortcoming that has been noted in prior work [31]. The same trend is observable in relation to the total number of occluded joints, see Fig. 7 b, and is clearly exacerbated as the majority of joints become occluded.

6. Conclusion

Robust uncertainty estimates must be a prerequisite for downstream decision-making, automated or otherwise, to avoid costly mistakes. We have shown two applications

of simultaneous aleatoric and epistemic uncertainty quantification in an HPE framework, focusing on the recently proposed DER approach, its plausibility of uncertainty estimates, and its relevance to the HPE domain. With added recalibration, DER is a promising contender for uncertainty quantification in HPE that does not rely on a costly resampling step. The fact of its portability will be of interest to real-world deployment on resource-constrained platforms as well as broader environmental considerations. While conceivable pitfalls, such as the disentanglement hinging on a single parameter estimate, need to be investigated further, the resultant uncertainties faithfully identify challenging situations in HPE. As such, the model generates theoretically (calibrated) and semantically (experiments) sound uncertainties, opening new opportunities for active learning and other downstream applications.

Acknowledgments

Page one photo credit https://pexels.com/@blue-bird: http://bitly.ws/BjFa

This research was supported by the Continental AG and the HEIDI project. HEIDI has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101069538. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 2, 3, 4, 11, 12

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 5, 6, 7

[3] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 2

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2

[5] Dennis Burgermeister and Cristóbal Curio. Pedrecnet: Multi-task deep neural network for full 3d human pose and orientation estimation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 441–448. IEEE, 2022. 6, 12

[6] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021. 2

[7] John Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3, 1990. 2

[8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2

[9] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015. 2

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 6

[11] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011. 2

[12] Nitesh B Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, volume 2, page 2, 2019. 3

[13] Chuchu Han, Xin Yu, Changxin Gao, Nong Sang, and Yi Yang. Single image based 3d human pose estimation via uncertainty learning. *Pattern Recognition*, 132:108934, 2022. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[15] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667, 2018. 2, 3

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5, 6, 7

[17] Moksh Jain, Salem Lahlou, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021. 2

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2, 3

[19] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015. 2

[20] Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pages 11683–11693. PMLR, 2022. 5

[21] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 2018. 5

[22] Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20448–20459, 2022. 3

[23] Yuandu Lai, Yucheng Shi, Yahong Han, Yunfeng Shao, Meiyu Qi, and Bingshuai Li. Exploring uncertainty in deep learning for construction of prediction intervals. *arXiv preprint arXiv:2104.12953*, 2021. 2

[24] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021. 3, 5, 6, 7

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5, 6, 7

[26] Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[28] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 3, 6

[29] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 2

[30] KV Mardia, JT Kent, and JM Bibby. Multivariate analysis, 1979. *Probability and mathematical statistics. Academic Press Inc*, 1979. 5

[31] Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9134–9142, 2023. 8

[32] Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2021. 5, 12

[33] Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007. 5, 11

[34] Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. Epistemic uncertainty sampling. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pages 72–86. Springer, 2019. 2

[35] Wendy S Parker. Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, 4(3):213–223, 2013. 2

[36] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR, 2018. 2

[37] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022. 14

[38] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 2

[39] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008. 2

[40] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019. 6

[41] Daniel Straub and Armen Der Kiureghian. Bayesian network enhanced with structural reliability methods: methodology. *Journal of engineering mechanics*, 136(10):1248–1258, 2010. 2

[42] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Computer Vision – ECCV 2018*, Lecture notes in computer science, pages 536–553. Springer International Publishing, Cham, 2018. 2, 3, 6

[43] Kai Wang, Michael Dohopolski, Qiongwen Zhang, David Sher, and Jing Wang. Towards reliable head and neck cancers locoregional recurrence prediction using delta-radiomics and learning with rejection option. *Medical physics*, 50(4):2212–2223, 2023. 2

[44] Xi Wang and Laurence Aitchison. Bayesian ood detection with aleatoric uncertainty and outlier exposure. *arXiv preprint arXiv:2102.12959*, 2021. 2

[45] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6, 7

[46] Yuejian Zhu. Ensemble forecast: A new approach to uncertainty and predictability. *Advances in atmospheric sciences*, 22(6):781–788, 2005. 2