MonoLoco: Monocular 3D Pedestrian Localization and Uncertainty Estimation

Lorenzo Bertoni, Sven Kreiss, Alexandre Alahi
EPFL VITA lab
CH-1015 Lausanne
lorenzo.bertoni@epfl.ch

Abstract

We tackle the fundamentally ill-posed problem of 3D human localization from monocular RGB images. Driven by the limitation of neural networks outputting point estimates, we address the ambiguity in the task by predicting confidence intervals through a loss function based on the Laplace distribution. Our architecture is a light-weight feed-forward neural network that predicts 3D locations and corresponding confidence intervals given 2D human poses. The design is particularly well suited for small training data, cross-dataset generalization, and real-time applications. Our experiments show that we (i) outperform state-of-the-art results on KITTI and nuScenes datasets, (ii) even outperform a stereo-based method for far-away pedestrians, and (iii) estimate meaningful confidence intervals. We further share insights on our model of uncertainty in cases of limited observations and out-of-distribution samples.

1. Introduction

Autonomous driving vehicles commonly rely on LiDAR sensing solutions despite high cost and sparsity of point clouds over long ranges [10, 59, 45]. Cost-effective perception systems have been proposed by adopting stereo/multiple cameras to address the fundamental ambiguity of monocular solutions [9, 32]. Yet researchers are studying how to push the limits of monocular perception to further contribute to multi-sensor fusion [33]. Progress has been made estimating 3D positions of vehicles from monocular images [8, 38, 48], while pedestrians have received far less attention due to lack of adequate performances. In fact, inferring 3D locations of pedestrians from a single image is particularly ambiguous due to the variance in human heights and shapes. In this work, we explicitly study the intrinsic ambiguity of locating pedestrians in the scene and investigate whether we can learn this ambiguity from the data. Driven by this perception task, we aim at providing more insights to the general problem of uncertainty estimation in deep learning.

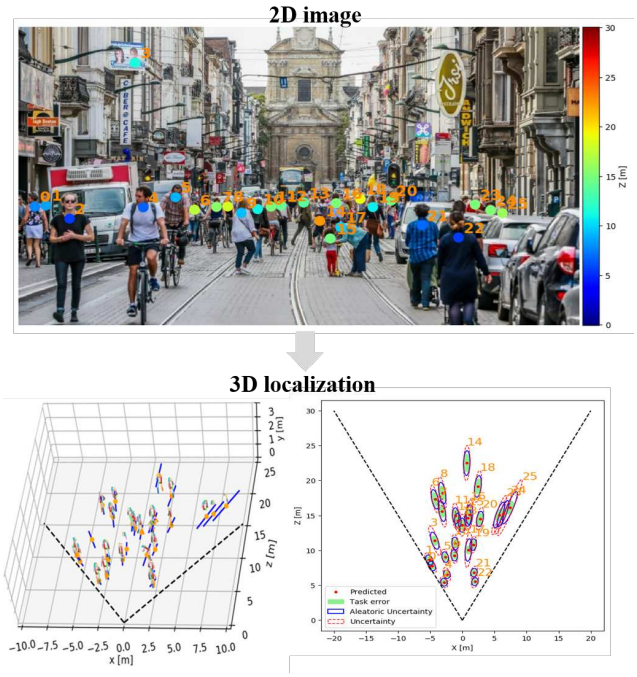

Figure 1. 3D localization of pedestrians from a single RGB image. Our method leverages 2D poses to find 3D locations as well as confidence intervals. The confidence intervals are shown as blue lines in the left 3D view and as ellipses in the right birds-eye-view.

Kendall and Gal [25] introduced practical uncertainty estimation for deep learning in perception tasks, distinguishing between *aleatoric* and *epistemic* uncertainty [11, 25]. The former models noise inherent in the observations, while the latter is a property of the model parameters and can be reduced by collecting more data. While their proposed measure of uncertainty is inspiring, they could not compare it with a known uncertainty, referred to as a *task error*. In this work, based on the statistical variation of human height within the adult population [52], we quantify the ambiguity of the task, *i.e.*, the task error: an upper bound of performances for monocular 3D pedestrian localization. Surprisingly, the task error is reasonably low. Our experiments

show accurate results in 3D localization without overcoming the limitation due to this intrinsic ambiguity.

We propose a simple probabilistic method for monocular 3D localization tailored for pedestrians. We specifically address the challenges of the ill-posed task by predicting confidence intervals in contrast to point estimates, which account for aleatoric and epistemic uncertainties. Our method is composed of two distinct steps. First, we leverage the exceptional progress of pose estimators to obtain 2D joints, a low-dimensional meaningful representation of humans. Second, we input the detected joints to a light-weight feed-forward network and output the 3D location of each instance along with a confidence interval. We explore whether 2D joints contain enough information for a network to learn the intrinsic ambiguity of the task as well as accurate localization. We leverage a recently introduced loss function based on the Laplace distribution [25] to incorporate aleatoric uncertainty for each predicted location without direct supervision at training time. MC dropout at inference time is used to capture epistemic uncertainty [16]. Our network, referred to as MonoLoco, independently learns the distribution of uncertainties, and predicts confidence intervals comparable with the corresponding task error. The code is publicly available online [1].

2. Related Work

Monocular 3D Object Detection. Recent approaches for monocular 3D object detection in the transportation domain focused only on vehicles as they are rigid objects with known shape. To the best of our knowledge, no previous work explicitly evaluated pedestrians from monocular RGB images. Kundegorski and Breckon [29] achieved reasonable performances combining infrared imagery and real-time photogrammetry. Alahi *et al*. combined monocular images with wireless signals [3] or with additional visual priors [1, 2]. The seminal work of Mono3D [8] exploited deep learning to create 3D object proposals for *car*, *pedestrian* and *cyclist* categories but it did not evaluate 3D localization of pedestrians. It assumed a fixed ground plane orthogonal to the camera and the proposals were then scored based on scene priors, such as shape, semantic and instance segmentations. Following methods continued to leverage Convolutional Neural Networks and focused only on *Car* instances. To regress 3D pose parameters from 2D detections, Deep3DBox [38], MonoGRnet [46], and Hu *et al*. [23] used geometrical reasoning for 3D localization, while Multi-fusion [57] and ROI-10D [35] incorporated a module for depth estimation. Recently, Roddick *et al*. [48] escaped the image domain by mapping image-based features into a birds-eye view representation using integral images. Another line of work fits 3D templates of cars to the image [54, 55, 7, 30].

While many of the related methods achieved reasonable performances for vehicles, current literature lacks monocular methods addressing other categories in the context of autonomous driving, such as pedestrians and cyclists.

Uncertainty in Computer Vision. Deep neural networks need to have the ability not only to provide the correct outputs but also a measure of uncertainty, especially in safety-critical scenarios like autonomous driving. Traditionally, Bayesian Neural Networks [47, 40] were used to model epistemic uncertainty through probability distributions over the model parameters. However, these distributions are often intractable and researchers have proposed interesting solutions to perform approximate Bayesian inference to measure uncertainty, including Variational Inference [20, 4, 50] and Deep Ensembles [31]. Alternatively, Gal *et al*. [16, 17] showed that applying dropout [51] at inference time yields a form of variational inference where parameters of the network are modeled as a mixture of multivariate Gaussian distributions with small variances. This technique, called Monte Carlo (MC) dropout, became popular also due to its adaptability to non-probabilistic deep learning frameworks.

In computer vision, uncertainty estimation using MC dropout has been applied for depth regression tasks [25], scene segmentation [39, 25] and, more recently, LiDAR 3D object detection for cars [14].

Human pose estimation. Detecting people in images and estimating their skeleton is a widely studied problem. State-of-the-art methods are based on Convolutional Neural Networks and can be grouped into top-down [43, 13, 21, 56] and bottom-up methods [6, 41, 42, 27].

Related to our work is Simple Baseline [36], which showed the effectiveness of latent information contained in 2D joints stimuli. They achieved state-of-the-art results by simply predicting 3D joints from 2D poses through a light, fully connected network. However, similarly to [37, 58, 49], they estimated relative 3D joint positions, not providing any information about the real 3D location in the scene.

3. Localization Ambiguity

Inferring depth of pedestrians from monocular images is a fundamentally ill-posed problem. This additional challenge is due to human variation of height. If every pedestrian has the same height, there would be no ambiguity. In this section, we quantify the ambiguity and analyze the maximum accuracy expected from monocular 3D pedestrian localization.

In our distance estimates, we assume that all humans have the same height h_{mean} and we analyze the error of this assumption. Inspired by Kundegorski and Breckon [29], we model the localization error due to variation of height

[1] **https://github.com/vita-epfl/monoloco**

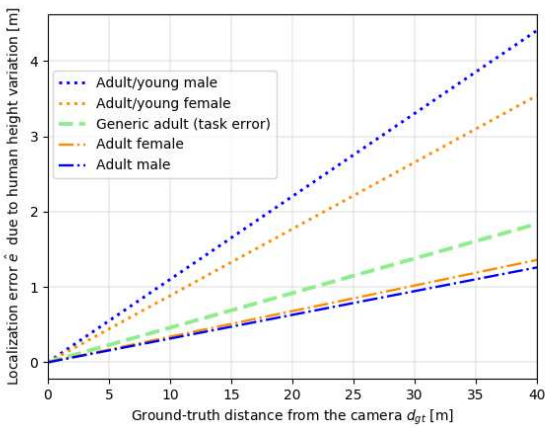

Figure 2. Localization error due to human height variations at different distances from the camera. We approximate the distribution of height for a generic adult as Gaussian mixture distribution and we define the *task error*: an upper bound of performances for monocular methods.

as a function of the ground truth distance from the camera, which we call *task error*. From the triangle similarity relation of human heights and distances, $d_{\text{h-mean}}/h_{\text{mean}} = d_{gt}/h_{gt}$, where h_{gt} and d_{gt} are the the ground-truth human height and distance, h_{mean} is the assumed mean height of a person and $d_{\text{h-mean}}$ the estimated distance under the h_{mean} assumption. We can define the task error for any person instance in the dataset as:

$$e \equiv |d_{gt} - d_{\text{h-mean}}| = d_{gt} \left| 1 - \frac{h_{\text{mean}}}{h_{gt}} \right| \quad . \tag{1}$$

Previous studies from a population of 63,000 European adults have shown that the average height is $178cm$ for males and $165cm$ for females with a standard deviation of around $7cm$ in both cases [52]. However, a pose detector does not distinguish between genders. Assuming that the distribution of human stature follows a Gaussian distribution for male and female populations [15], we define the combined distribution of human heights, a Gaussian mixture distribution $P(H)$, as our unknown ground-truth height distribution. The *expected task error* becomes

$$\hat{e} = d_{gt}\ E_{h \sim P(H)} \left[\left| 1 - \frac{h_{mean}}{h} \right| \right] \tag{2}$$

which represents a lower bound for monocular 3D pedestrian localization due to the intrinsic ambiguity of the task. The analysis can be extended beyond adults. A 14-year old male reaches about 90% of his full height and a female about 95% [15, 29]. Including people down to 14 years old leads to an additional source of height variation of 7.9% and 5.6% for men and women, respectively [29]. Figure 2 shows the expected localization error $\hat{e}$ due to height variations in different cases as a function of the ground-truth distance from the camera d_{gt}. This analysis shows that the ill-posed problem of localizing pedestrians, while imposing an intrinsic limit, does not prevent from robust localization in general cases.

4. Method

The goal of our method is to detect pedestrians in 3D given a single image. We argue that effective monocular localization implies not only accurate estimates of the distance but also realistic predictions of uncertainty. Consequently, we propose a method which learns the ambiguity from the data without supervision and predicts confidence intervals in contrast to point estimates. The task error modeled in Eq. 2 allows to compare the predicted confidence intervals with the intrinsic ambiguity of the task.

Figure 3 illustrates our overall method, which consists of two main steps. First, we exploit a pose detector to escape the image domain and reduce the input dimensionality. 2D joints are a meaningful low-level representation which provides invariance to many factors, including background scenes, lighting, textures and clothes. Second, we use the 2D joints as input to a feed-forward neural network which predicts the distance and the associated ambiguity of each pedestrian. In the training phase, there is no supervision for the localization ambiguity. The network implicitly learns it from the data distribution.

4.1. Setup

Input. We use a pose estimator to detect a set of keypoints $[u_i, v_i]^T$ for every instance in the image. We then back-project each keypoint i into normalized image coordinates $[x_i^*, y_i^*, 1]^T$ using the camera intrinsic matrix K:

$$[x_i^*, y_i^*, 1]^T = K^{-1} [u_i, v_i, 1]^T . \tag{3}$$

This transformation is essential to prevent the method from overfitting to a specific camera. Furthermore, even if we are not predicting a relative 3D location but the distance to the camera, we zero-center the 2D inputs around the center coordinates. This ensures that the model uses relative distances between joints to make predictions and it prevents overfitting on specific locations in the image.

2D Human Poses. We obtain 2D joint locations of pedestrians using two off-the-shelf pose detectors: the top-down method Mask R-CNN [21] and the bottom-up one PifPaf [28], both trained on the COCO dataset [34]. The detector can be regarded as a stand-alone module independent from our network, which uses 2D joints as inputs. None of the detectors has been fine-tuned on KITTI or nuScenes datasets as no annotations for 2D poses are available.

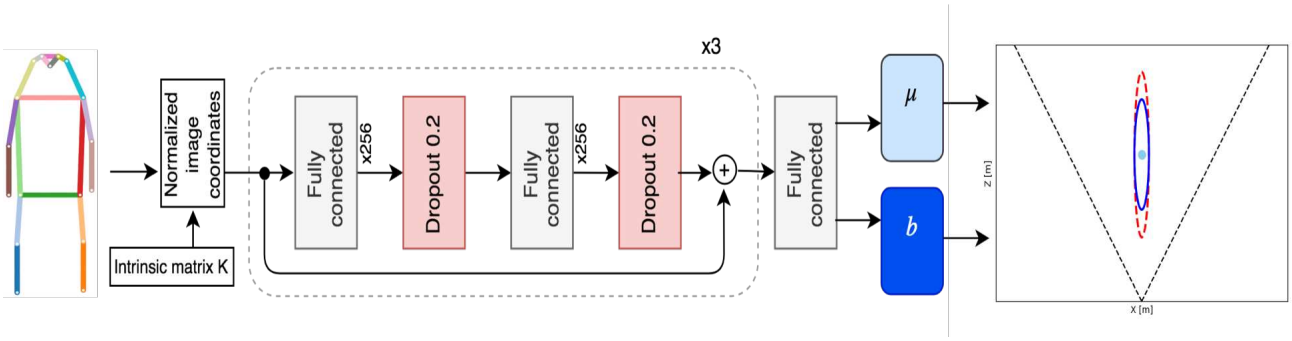

Figure 3. Network architecture. The input is a set of 2D joints extracted from a raw image and the output is the 3D location of a pedestrian μ and the spread b which represents the associated *aleatoric* uncertainty. The confidence interval is obtained as $\mu \pm b$. *Epistemic* uncertainty is obtained through stochastic forward passes applying MC dropout [16]. The dashed ellipse represents the two combined uncertainties. Every fully connected layer outputs 256 features and is followed by a Batch Normalization layer [24] and a ReLU activation function.

Output. We parametrize the 3D physical location of each instance through its center location $\mathbf{D} = [x_c, y_c, z_c]^T$. We further assume that the projection of the center into the image plane corresponds to the center of the detected bounding box $[u_c, v_c]^T$. Under these settings, the location of each pedestrian has three degrees of freedom and two constraints. We choose to regress the norm of the vector $||\mathbf{D}||_2 = \sqrt{x_c^2 + y_c^2 + z_c^2}$ to further constrain the location of a pedestrian. For brevity, we will use the notation $d = ||\mathbf{D}||_2$. The main criterion is that the dimensions of any object projected into the image plane only depend on the norm of the vector $\mathbf{D}$ and they are not affected by the combination of its components. The same pedestrian in front of a camera or at the margin of the camera field-of-view will appear as having the same height in the image plane, as long as the distance from the camera d is the same.

Base Network. The building blocks of our model are shown in Figure 3. The architecture, inspired by Martinez *et al*. [36], is a simple, deep, fully-connected network with six linear layers with 256 output features. It includes dropout [51] after every fully connected layer, batch-normalization [24] and residual connections [22]. The model contains approximately 400k training parameters.

4.2. Uncertainty

In this work, we propose a probabilistic network which models two types of uncertainty: *aleatoric* and *epistemic* [11, 25].

Aleatoric uncertainty is an intrinsic property of the task and the inputs. It does not decrease when collecting more data. In the context of 3D monocular localization, the intrinsic ambiguity of the task represents a quota of aleatoric uncertainty. In addition, some inputs may be more noisy than others, leading to an input-dependent aleatoric uncertainty. Epistemic uncertainty is a property of the model parameters, and it can be reduced by gathering more data. It is useful to quantify the ignorance of the model about the collected data, *e.g.*, in case of out-of-distribution samples.

Aleatoric Uncertainty. Aleatoric uncertainty is captured through a probability distribution over the model outputs. We define a relative Laplace loss based on the negative log-likelihood of a Laplace distribution as:

$$L_{\text{Laplace}}(x|\mu, b) = \frac{|1 - \mu/x|}{b} + \log(2b) \qquad (4)$$

where x is the ground truth and $\{\mu, b\}$ are the parameters predicted by the model. μ represents the predicted distance while b is the spread, making this training objective an attenuated L_1-type loss via spread b. During training, the model has the freedom to predict a large spread b, leading to attenuated gradients for noisy data. At inference time, the model predicts the distance μ and a spread b which indicates its confidence about the predicted distance. Following [25], to avoid the singularity for $b = 0$, we apply a change of variable to predict the log of the spread $s = \log(b)$.

Compared to previous methods [25, 53], we design a Laplace loss which works with relative distances to keep into account the role of distance in our predictions. Estimating the distance of a pedestrian with an absolute error can lead to a fatal accident if the person is very close, or be negligible if the same human is far away from the camera.

Epistemic Uncertainty. To model epistemic uncertainty, we follow [16, 25] and consider each parameter as a mixture of two multivariate Gaussians with small variances and means 0 and θ. The additional minimization objective for N data points is:

$$L_{\text{dropout}}(\theta, p_{drop}) = \frac{1 - p_{drop}}{2N} ||\theta||^2 \quad . \qquad (5)$$

In practice, we perform dropout variational inference by training the model with dropout before every weight layer and then performing a series of stochastic forward passes at

Method	Type	ALP [%]			ALE [m]		
		$< 0.5m$	$< 1m$	$< 2m$	$Easy$	$Moderate$	$Hard$
Mono3D [8]	Mono	13.2	23.2	38.9	2.13 (2.32)	2.85 (3.09)	3.68 (4.46)
MonoDepth [19] + PifPaf [28]	Mono	20.5	35.3	50.6	1.48 (1.69)	2.32 (2.99)	3.03 (3.67)
Our Geometric baseline	Mono	16.6	32.6	62.2	1.40 (1.48)	1.35 (1.69)	1.61 (1.91)
Our MonoLoco - trained on KITTI	Mono	29.0	49.6	71.2	0.94 (0.98)	1.09 (1.49)	1.27 (1.90)
Our MonoLoco - trained on nuScenes	Mono	**30.8**	**51.7**	**72.1**	**0.86 (0.92)**	**1.00 (1.25)**	**1.17 (1.65)**
3DOP [9]	Stereo	41.4	54.9	63.2	0.63 (0.71)	1.18 (1.27)	1.94 (2.11)
Task Error	-	49.0	67.3	80.0	0.62 (0.55)	0.68 (0.99)	0.64 (0.75)

Table 1. Comparing our proposed method against baseline results on KITTI dataset [18]. The ALE metric is reported for pedestrians commonly detected by all methods to make fair comparison and, on parenthesis, for all the pedestrians detected by each method independently. We outperform all monocular methods and we achieve comparable performances against 3DOP which leverages stereo images for training and testing. Our method uses monocular images and shows cross-dataset generalization when trained on nuScenes dataset [5]. We use PifPaf [28] as off-the-shelf network to extract 2D poses.

test time using the same dropout probability p_{drop} of training time. The use of fully-connected layers makes the network particularly suitable for this approach, which does not require any substantial modification of the model.

The combined epistemic and aleatoric uncertainties are captured by the sample variance of predicted distances $\tilde{x}$. They are sampled from multiple Laplace distributions parameterized with the predictive distance μ and spread b from multiple forward passes with MC dropout:

$$Var(\tilde{X}) = \frac{1}{TI}\sum_{t=1}^{T}\sum_{i=1}^{I}\tilde{x}_{t,i}^2(\mu_t, b_t) - \left[\frac{1}{TI}\sum_{t=1}^{T}\sum_{i=1}^{I}\tilde{x}_{t,i}(\mu_t, b_t)\right]^2 \quad (6)$$

where for each of the T computationally expensive forward passes, I computationally cheap samples are drawn from the Laplace distribution.

5. Experiments

5.1. Implementation details.

Datasets. We train and evaluate our model on KITTI Dataset [18]. It contains 7481 training images along with camera calibration files. All the images are captured in the same city from the same camera. To analyze cross-dataset generalization properties, we train another model on the teaser of the recently released *nuScenes* dataset [5] and we test it on KITTI. We do not perform cross-dataset training.

Training/Evaluation Procedure. To obtain input-output pairs of 2D joints and distances, we apply an off-the-shelf pose detector and use intersection over union of 0.3 to match our detections with the ground-truths, obtaining 5000 instances for KITTI and 14500 for nuScenes teaser. KITTI images are upscaled by a factor of two to match the minimum dimension of 32 pixels of COCO instances. NuScenes already contains high-definition images, which are not modified. We follow the KITTI train/val split of Chen *et al*. [8] and we run the training procedure for 200 epochs using Adam optimizer [26], a learning rate of 10^{-3} and mini-batches of 512. The code, available online [1], is developed using PyTorch [44]. Working with a low-dimensional latent representation is very appealing as it allows fast experiments with different architectures and hyperparameters.

5.2. Evaluation.

Localization Error. We evaluate 3D pedestrian localization using the Average Localization Precision (ALP) metric defined by Xiang *et al*. [54] for the *car* category. ALP considers a prediction as correct if the error between the predicted distance and the ground-truth is smaller than a threshold. We also analyze the average localization error (ALE) in two different conditions. Following KITTI guidelines, we split the instances in three difficulty regimes based on bounding box height, levels of occlusion and truncation: *easy*, *medium* and *hard*. We also compare the results against the task error of Eq. 2, which defines the target error for monocular approaches due to the ambiguity of the task.

Geometrical Approach. 3D pedestrian localization is an ill-posed task due to human height variations. On the other side, estimating the distance of an object of known dimensions from its projections into the image plane is a well-known deterministic problem. As a baseline, we consider humans as fixed objects with the same height and we investigate the localization accuracy under this assumption.

For every pedestrian, we apply a pose detector to calculate distances in pixels between different body parts in the image domain. Combining this information with the location of the person in the world domain, we analyze the distribution of the real dimensions (in meters) of all the instances in the training set for three segments: head to shoulder, shoulder to hip and hip to ankle.

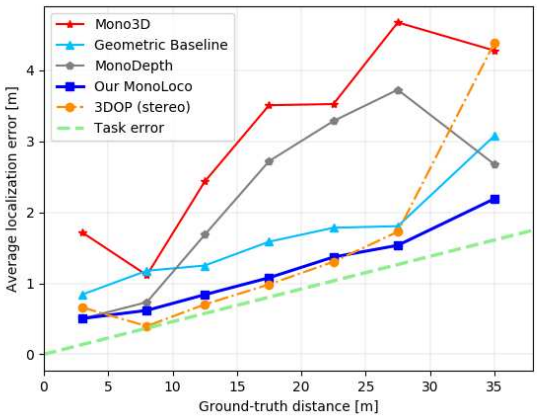

Figure 4. Average localization error for the instances commonly detected by all methods. We outperform the monocular Mono3D [8] while achieving comparable performances with the stereo 3DOP [9]. Monocular performances are bounded by our modeled task error in Eq. 2.

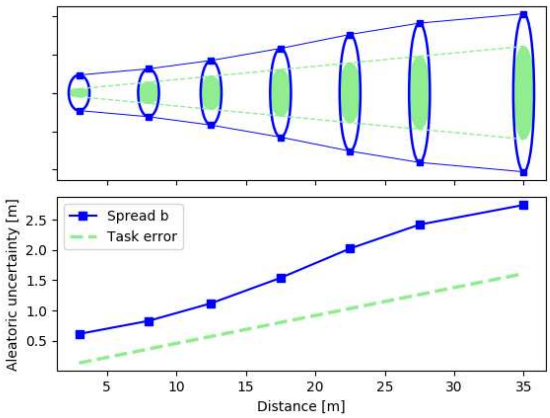

Figure 5. Results of aleatoric uncertainty predicted by MonoLoco (spread b) and the modeled aleatoric uncertainty due to human height variation (task error $\hat{e}$). The term $b - \hat{e}$ is indicative of the aleatoric uncertainty due to noisy observations. On the top figure, we visualize the average predicted and ground truth confidence intervals $\pm b$ and $\pm \hat{e}$ at various distances, using ellipses with minor axis of one meter as a reference.

	$\lvert x-\mu\rvert/\sigma$	$\lvert\sigma - e\rvert$ [m]	Recall [%]
$p_{drop} = 0.05$	0.60	0.90	82.8
$p_{drop} = 0.2$	0.58	0.96	84.3
$p_{drop} = 0.4$	0.50	1.26	88.3

Table 2. Precision and recall of uncertainty for KITTI validation set with 50 stochastic forward passes. $|x - \mu|$ is the localization error, σ the predicted confidence interval, $\hat{e}$ the task error modeled in Eq. 2 and Recall is represented by the % of ground truth instances inside the predicted confidence interval.

For our calculation we assume a pinhole model of the camera and that all instances stand upright. Using the camera intrinsic matrix K and knowing the ground truth location of each instance $\mathbf{D} = [x_c, y_c, z_c]^T$ we can back-project each keypoint from the image plane to its 3D location and measure the height of each segment using Eq. 3. We calculate the mean and the standard deviation in meters of each of the segments for all the instances in the training set. The standard deviation is used to choose the most stable segment for our calculations. For instance, the position of the head with respect to shoulders may vary a lot for each instance. To take into account noise in the 2D joints predictions we also average between left and right keypoints values. The result is a single height Δy_{1-2} which represents the average length of two body parts. In practice, our geometric baseline uses the *shoulder-hip* segment and predicts an average height of $50.5cm$. Combining the study on human heights [52] described in Section 3 with the anthropometry study of Drillis *et al*. [12], we can compare our estimated Δy_{1-2} with the human average *shoulder-hip* height: $0.288 * 171.5cm = 49.3cm$.

The next step is to calculate the location of each instance knowing the value in pixels of the chosen keypoints v_1 and v_2 and assuming Δy_{1-2} to be their relative distance in meters. This configuration requires to solve an over-constrained linear system with two specular solutions, of which only one is inside the camera field of view.

Baselines. We compare our method on KITTI against two monocular approaches and a stereo one:

- **Mono3D** [8] is a monocular 3D object detector for cars, cyclists and pedestrians. 3D localization of pedestrians is not evaluated but detection results are publicly available.
- **MonoDepth** [19] is a monocular depth estimator which predicts a depth value for each pixel in the image. To estimate a reference depth value for every pedestrian, we detect 2D joints using PifPaf and calculate the depth for a set of 9 pixels around each keypoint. We then consider the minimum depth as our reference value. Experimentally, the minimum depth increases the performances compared to the average one. From the depth, we calculate the distance d using the normalized image coordinates of the center of the bounding box.
- **3DOP** [9] is a stereo approach for pedestrians, cars and cyclists and their 3D detections are publicly available.

5.3. Results.

Localization Accuracy. Table 1 summarizes our quantitative results on KITTI. We strongly outperform all the other monocular approaches on all metrics with any of the

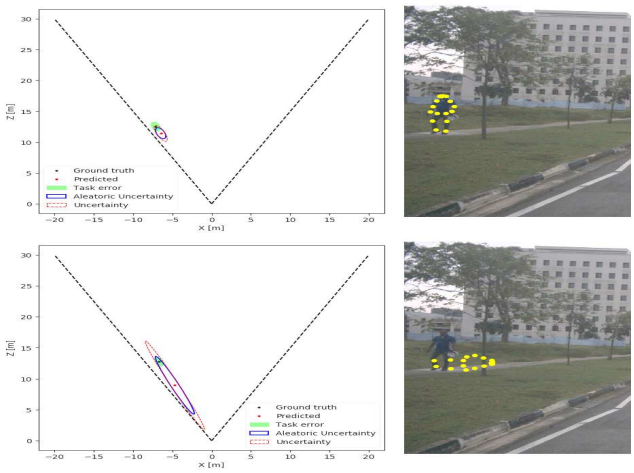

Figure 6. Simulating the outlier case of a person lying on the ground. In the top image, the predicted confidence interval is small and the detection accurate. In the bottom image, we create an outlier pose by projecting on the ground the original pose. The network predicts higher uncertainty, a useful indicator to warn about out-of-distribution samples.

Mask R-CNN [21]	ALE [m] $^{10}_{0}$	$^{20}_{10}$	$^{30}_{20}$	$^{+}_{30}$	*All*
Geometric	0.79	1.52	3.17	9.08	3.73
L_1 loss	0.85	**1.17**	2.24	4.11	2.14
Gaussian loss	0.90	1.28	2.34	4.32	2.26
Laplace Loss	**0.74**	**1.17**	2.25	4.12	2.12
PifPaf [28]	ALE [m] $^{10}_{0}$	$^{20}_{10}$	$^{30}_{20}$	$^{+}_{30}$	*All*
Geometric	0.83	1.40	2.15	3.59	2.05
L_1 loss	0.83	1.24	**2.09**	3.32	1.92
Gaussian loss	0.89	1.22	2.14	3.50	1.97
Laplace Loss	0.75	1.19	2.24	**3.25**	**1.90**

Table 3. Impact of different loss functions and pose detectors on nuScenes teaser validation set [5].

Method \ Time [ms]	t^{pose}	t^{model}	t^{total}
Mono3D [8]	-	1800	1800
3DOP [9]	-	2000	2000
Our (1 forward pass)	89 / 162	10	**99 / 172**
Our (50 forward passes)	89 / 162	51	140 / 213

Table 4. Single-image inference time on a GTX 1080Ti for KITTI dataset with Pifpaf as pose detector. We only considered images with positive detections. Most computation comes from the pose detector (ResNet 50 / ResNet 152 backbones). For Mono3D and 3DOP we report published statistics on a Titan X GPU.

two models trained either on KITTI or nuScenes. We obtain comparable results with the stereo approach 3DOP [9], which has been trained and evaluated on KITTI and makes use of stereo images during training and test time.

In Figure 4, we make an in-depth comparison analyzing the average localization error as a function of the ground truth distance, while Figure 7 shows qualitative results on challenging images from KITTI and nuScenes datasets. A video with additional results is available online. [2].

Aleatoric Uncertainty. We compare in Figure 5 the aleatoric uncertainty predicted by our network through spread b with the *task error* due to human height variation defined in Eq. 2. The predicted spread b is a property of each set of inputs and, differently from $\hat{e}$, is not only a function of the distance from the camera d. Indeed, the predicted aleatoric uncertainty includes not only the uncertainty due to the ambiguity of the task, but also the uncertainty due to noisy observations [25], *i.e.*, the 2D joints inferred by the pose detector. Hence, we can approximately define the predictive aleatoric uncertainty due to noisy joints as $b - \hat{e}$ and we observe that the further a person is from the camera, the higher is the term $b - \hat{e}$. The spread b is the result of a probabilistic interpretation of the model and the resulting confidence intervals are calibrated. On KITTI validation set they include 68% of the instances.

Combined Uncertainty. The combined aleatoric and epistemic uncertainties are captured by sampling from multiple Laplace distributions using MC dropout. The magnitude of the uncertainty depends on the chosen dropout probability p_{drop} in Eq. 5. In Table 2, we analyze the precision/recall trade-off for different dropout probabilities and choose $p_{\text{drop}} = 0.2$. We perform 50 computationally expensive forward passes and, for each of them, 100 computationally cheap samples from Laplace distribution using Eq. 6. As a result, 84% of pedestrians lie inside the predicted confidence intervals for the validation set of KITTI.

Our final goal is to make self-driving cars safe and being able to predict a confidence interval instead of a single regression number is a first step towards this direction. To illustrate the benefits of predicting intervals over point estimates, we construct a controlled risk analysis. We define as *high-risk cases* all those instances where the ground truth distance is smaller than the predicted one, hence a collision is more likely to happen. We estimate that among the 1932 detected pedestrians in KITTI which match a ground truth, 48% of them are considered as *high-risk cases*, but for 89% of them the ground truth lies inside the predicted interval.

Outliers. Leveraging on the simplicity of manipulation of 2D joints, we analyze the role of the predicted uncertainty in case of an outlier. As shown in Figure 6, we recreate the pose of a person lying down and we compare it with a "standard" detection of the same person standing up. When the pedestrian is lying down, the network predicts an unusually large confidence interval which includes the ground truth location.

[2] `https://youtu.be/ii0fqerQrec`

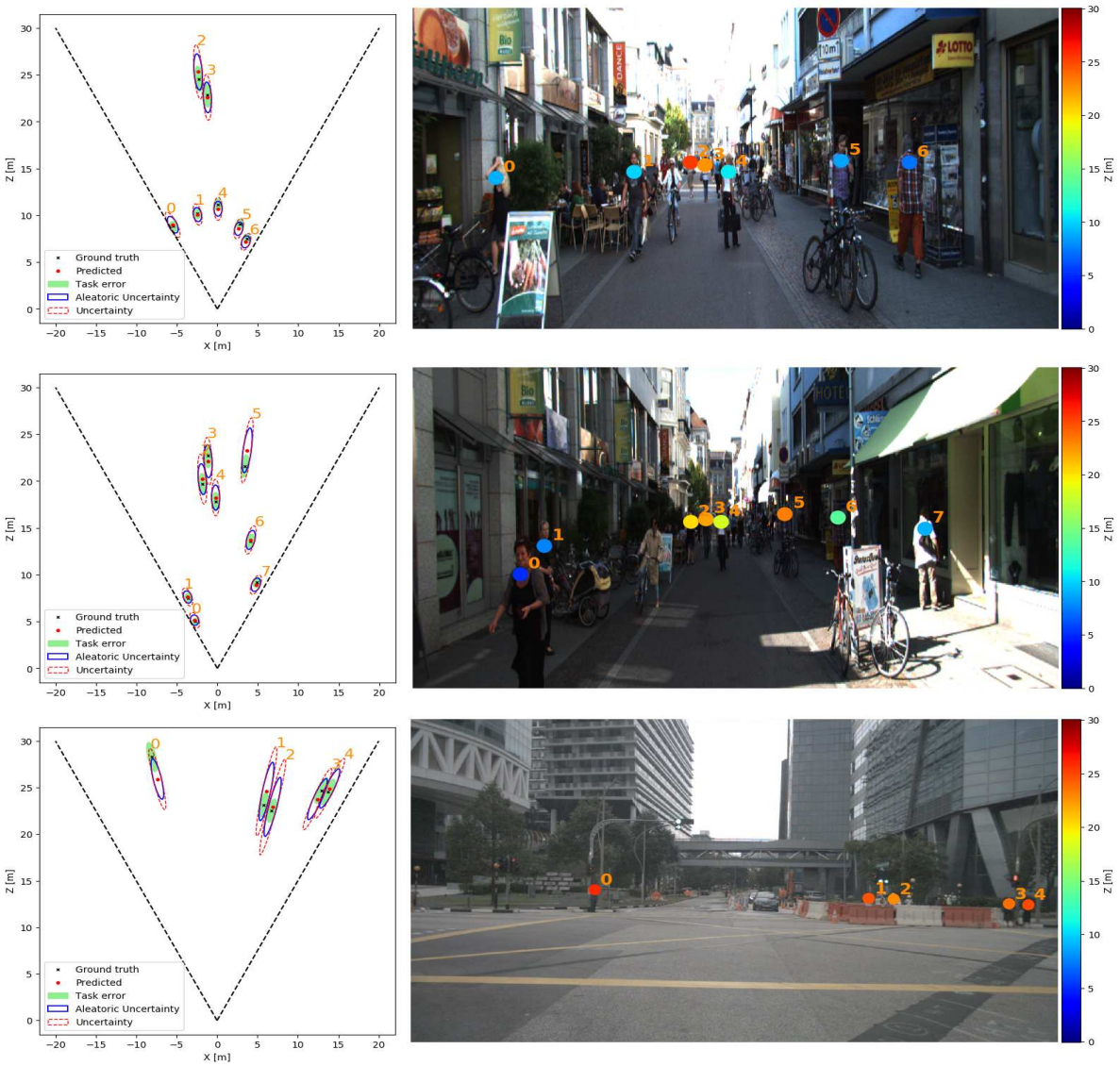

Figure 7. Illustration of results from KITTI [18] (top and middle) and nuScenes [5] (bottom) datasets containing true and inferred distance information as well as confidence intervals (represented by ellipses with minor axis of one meter). We observe that the predicted uncertainty increases in case of occlusions (bottom image, pedestrians 1 and 2).

In the bottom image of Figure 7, we also highlight the behavior of the model in case of partially occluded pedestrians (pedestrians 1 and 2), where we also empirically observe larger confidence intervals when compared to visible pedestrians at similar distances.

Ablation studies. In Table 3 we analyze the effects of choosing a top-down or a bottom-up pose detector with different loss functions and with our deterministic geometric baseline. L_1-type losses perform slightly better than the Gaussian loss, but the main improvement is given by choosing PifPaf as pose detector.

Run time. A run time comparison is shown in Table 4. Our method is 9-20 times faster than compared methods (depending on the pose detector backbone) and it is the only one suitable for real-time applications.

6. Conclusions

We have proposed a new approach for 3D pedestrian localization based on monocular images which addresses the intrinsic ambiguity of the task by predicting calibrated confidence intervals. We have shown that our method even outperforms a stereo approach at further distances because it is less sensitive to low-resolution imaging issues.

For autonomous driving applications, combining our method with a stereo approach is an exciting direction for accurate, low-cost and real-time 3D localization.

Acknowledgements We acknowledge the support of Samsung and Farshid Moussavi for helpful discussions.

References

[1] Alexandre Alahi, Michel Bierlaire, and Murat Kunt. Object detection and matching with mobile cameras collaborating with fixed cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.

[2] Alexandre Alahi, Michel Bierlaire, and Pierre Vandergheynst. Robust real-time pedestrians detection in urban environments with low-resolution cameras. *Transportation research part C: emerging technologies*, 39:113–128, 2014.

[3] Alexandre Alahi, Albert Haque, and Li Fei-Fei. Rgb-w: When vision meets wireless. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3289–3297, 2015.

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *the International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1613–1622. PMLR, 2015.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[7] Florian Chabot, Mohamed Ali Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, pages 1827–1836, 2017.

[8] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2147–2156, 2016.

[9] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015.

[10] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017.

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

[12] Rudolfs Drillis, Renato Contini, and Maurice Bluestein. *Body segment parameters*. New York University, School of Engineering and Science, 1969.

[13] Haoshu Fang, Shuqin Xie, and Cewu Lu. Rmpe: Regional multi-person pose estimation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017.

[14] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 3266–3273, 2018.

[15] JV Freeman, TJ Cole, S Chinn, PRh Jones, EM White, and MA Preece. Cross sectional stature and weight reference curves for the uk, 1990. *Archives of disease in childhood*, 73(1):17–24, 1995.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *the International Conference on Machine Learning*, pages 1050–1059, 2016.

[17] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017.

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[20] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[22] 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*, pages 770–778, 2016.

[23] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krhenbhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. *arXiv*, abs/1811.10742, 2018.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5574–5584, 2017.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–433, 2018.

[28] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11977–11986, 2019.

[29] Mikolaj E Kundegorski and Toby P Breckon. A photogrammetric approach for real-time 3d localization and tracking of pedestrians in monocular infrared imagery. In *SPIE Optics and Photonics for Counterterrorism, Crime Fighting, and Defence*, volume 9253, 2014.

[30] Abhijit Kundu, Yin Li, and James M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3559–3568, 2018.

[31] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[32] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7644–7652, 2019.

[33] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018.

[34] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[35] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2069–2078, 2019.

[36] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668. IEEE, 2017.

[37] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1561–1570, 2017.

[38] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7074–7082, 2017.

[39] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

[40] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[41] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017.

[42] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[43] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Christoph Bregler, and Kevin P. Murphy. Towards accurate multi-person pose estimation in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3711–3719, 2017.

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[45] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2018.

[46] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019.

[47] Michael D. Richard and Richard Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

[48] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *the British Machine Vision Conference (BMVC)*, 2019.

[49] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[50] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *the International Conference on Machine Learning*, pages 1218–1226, 2015.

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[52] Peter M Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489, 2008.

[53] Sascha Wirges, Marcel Reith-Braun, Martin Lauer, and Christoph Stiller. Capturing object detection uncertainty in multi-layer grid maps. *arXiv preprint arXiv:1901.11284*, 2019.

[54] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Data-driven 3d voxel patterns for object category recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1903–1911, 2015.

[55] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Subcategory-aware convolutional neural networks for object

proposals and detection. In *the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 924–933, 2017.

[56] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[57] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2353, 2018.

[58] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, pages 8410–8419, 2018.

[59] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018.