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 with a new neural network predicting confidence intervals through a loss function based on the Laplace distribution. Our architecture is a light-weight feed-forward neural network which predicts the 3D coordinates given 2D human pose. The design is particularly well suited for small training data and cross-dataset generalization. Our experiments show that (i) we outperform state-of-the art results on KITTI and nuScenes datasets, (ii) even outperform stereo for far-away pedestrians, and (iii) estimate meaningful confidence intervals. We further share insights on our model of uncertainty in case of limited observation and out-of-distribution samples.

1. Introduction

The complexity of monocular 3D localization can be attributed to the fundamental ambiguity of locating objects in the world based on their observed position in an image. This ambiguity is particularly relevant for pedestrians, which are characterized by different heights and shapes. To overcome this limitation, the majority of autonomous driving applications are based on LiDAR and sensor fusion, despite high cost and sparsity of point clouds over long ranges. In addition, datasets for 3D vision are smaller compared to traditional datasets for 2D vision tasks, as they require complex and expensive labeling procedures. Inferring 3D information of a scene from a single image remains a critical task yet to be solved to assure low-cost mobility. Progress has been made on estimating the 3D position of vehicles, while all other classes, including pedestrians, have received far less attention due to lack of adequate performances [39, 4].

Perception systems often do not take into account measures of confidence in their predictions. Uncertainty estimation is crucial in monocular 3D vision when dealing with the intrinsic ambiguity of locating 3D objects in the scene. It is also essential for deep learning systems to convey trust in autonomous driving applications, where human safety is at stake. Kendall and Gal [20] introduced practical uncertainty estimation for deep learning in computer vision, distinguishing between *aleatoric* and *epistemic* uncertainty [6, 20]. 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.

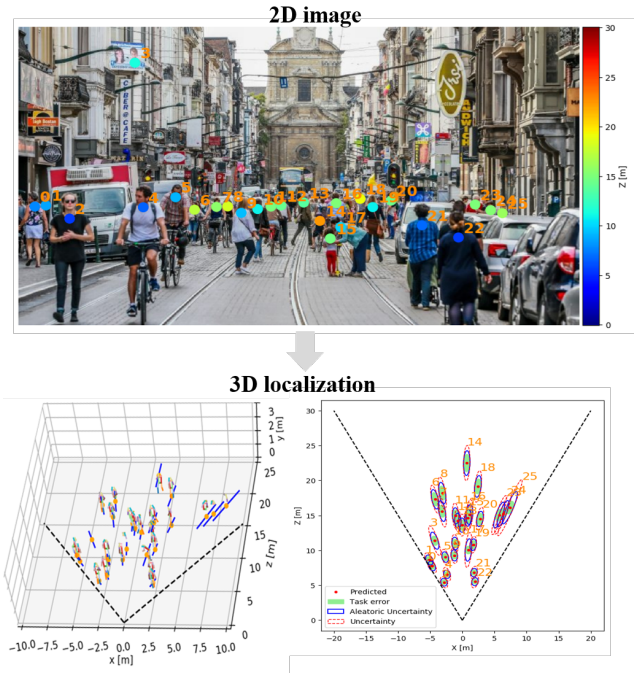

Figure 1. 3D localization of pedestrians from a single RGB image. Our method leverages 2D poses to find 3D locations of pedestrians (3D view on the left) as well as confidence intervals (birds-eye-view on the right) to address the ambiguity of the task.

In this work, 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 uncertain-

ties. 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 weather 2D joints contain enough information with the goal of learning the intrinsic ambiguity of the task as well as accurate localization. We leverage a recently introduced loss function based on the Laplace distribution [20] to incorporate aleatoric uncertainty for each predicted location without direct supervision at training time. MC Dropout at prediction time is used to capture epistemic uncertainty [10]. In parallel, based on the statistical variation of human height within the adult population [42], we quantify the ambiguity of the task, which we call *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. Furthermore, our network, referred to as MonoLoco, independently learns the distribution of uncertainties, predicting confidence intervals comparable with the corresponding task error. The code is open source and 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 [24] achieved reasonable performances combining infrared imagery and real-time photogrammetry. The seminal work of Mono3D [4] 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 on Convolutional Neural Networks and focused only on *Car* instances. To regress 3D pose parameters from 2D detections, Deep3DBox [30] and Hu *et al*. [17] leveraged on geometrical constraints for localization, while Multi-fusion [48] and ROI-10D [28] incorporated a module for depth estimation. Recently, Roddick *et al*. [39] 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 [45, 46, 3, 25].

While many of the related methods are achieving reasonable performances for cars, 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 [38, 32] 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 [14, 1, 40] and Deep Ensembles [26]. Alternatively, Gal *et al*. [10, 11] showed that applying dropout [41] 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 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 [20], scene segmentation [31, 20] and, more recently, LiDAR 3D object detection for cars [8].

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 [36, 7, 18, 15, 47] and bottom-up methods [2, 33, 35, 22].

Related to our work is Simple Baseline [29], 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 they estimated 3D joint positions relative to the hip joint, 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.

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 [42]. Following the approach of Kundegorski and Breckon [24], we model the localization

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

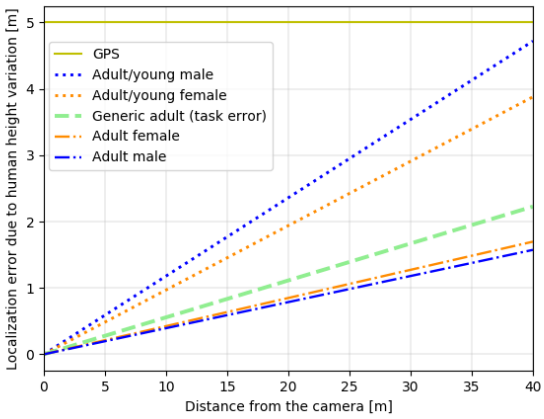

Figure 2. Localization error due to human height variations at different distances from the camera. We approximate the distribution of height for a generic adult as Gaussian mixture distribution with a standard deviation of $9.5cm$ and we define it as the *task error*: an upper bound of performances for monocular methods.

error due to a $7cm$ of standard deviation of height as a function of the distance from the camera. However, a pose detector does not distinguish between genders. Assuming that the distribution of human stature follows a Gaussian distribution [9], we estimate the combined localization error due to male and female height variations as a Gaussian mixture distribution. We obtain a distribution with standard deviation of $\sim 9.6cm$ and a relative error of 5.6%. We define the *task error* as:

$$e = m * d \tag{1}$$

where m is a fixed coefficient representing the relative error and d the distance from the camera. The *task error* e represents a lower bound for 3D pedestrian localization due to the intrinsic ambiguity of the task. The analysis can be extended beyond adults. A 14-year old male reaches 90% of his full height and a female about 95% [9]. 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 [24]. Figure 2 shows the localization error due to height variation in different cases as a function of the distance from the camera. We also include in the comparison the consumer-level Global Position System (GPS) accuracy, which is approximately $\pm 5m$ under ideal conditions [43, 24]. Even at 40 meters from the camera, the task error is smaller than the GPS one. 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 estimate 3D pedestrian localization in egocentric coordinates given monocular images. 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 and predicts confidence intervals in contrast to point estimates. The *task error* modeled in eq 1 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. 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 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:

$$[x_i^*, y_i^*, 1]^T = K^{-1} [u_i, v_i, 1]^T . \tag{2}$$

This transformation is essential to prevent the method from overfitting to a specific camera or dataset. 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 hip joint. This ensures that the model only uses relative distances between joints to make predictions and it prevents overfitting on specific locations of the image.

2D Human Poses. We obtain 2D joint locations of pedestrians using two state-of-the-art pose detectors: the top-down method MASK R-CNN [15] and the bottom-up one PifPaf [23], both trained on the COCO Dataset [27]. The detector can be regarded as a stand-alone module, where its predictions are the input of our network. None of the detectors have been fine-tuned on KITTI or nuScenes datasets as no annotations for 2D poses are available.

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 (DoF) and

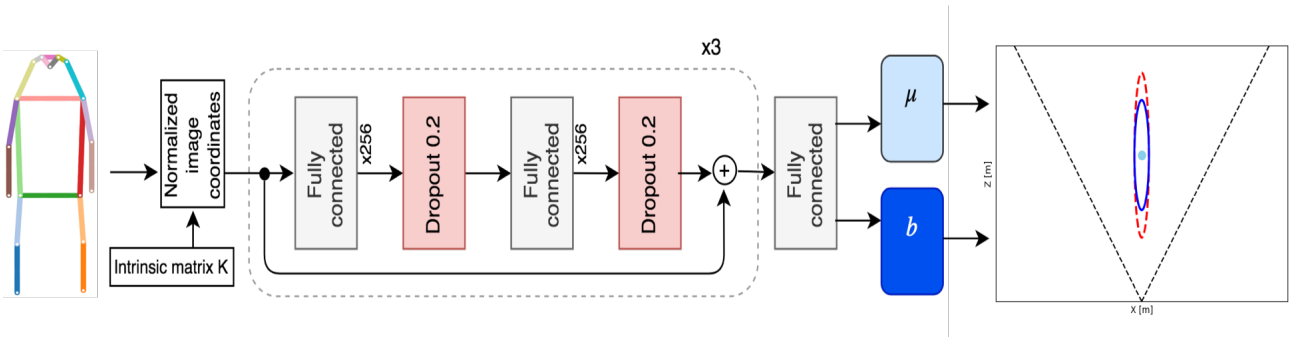

Figure 3. Network architecture. The input is a set of 2D joints extracted from a raw image and the output is the 3D location of a pedestrian μ and the spread b which represents the associated *aleatoric* uncertainty. The confidence interval is obtained as $\mu \pm b$. *Epistemic* uncertainty is obtained through stochastic forward passes applying dropout at test time. The dashed ellipse represents the two combined uncertainties. Every fully connected layer outputs 256 features and is followed by a Batch Normalization layer [19] and a ReLU activation function.

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*. [29], is a simple, deep, fully-connected network with six linear layers with 256 output features. It includes dropout [41] after every fully connected layer, batch-normalization [19] and residual connections [16]. The model contains approximately 400k training parameters.

4.2. Uncertainty

In this work, we propose a probabilistic network which models two types of uncertainties: *aleatoric* and *epistemic* [6, 20].

Aleatoric uncertainty is an intrinsic property of the task and the inputs and 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 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:

$$L_{\text{Laplace}}(x|\mu, b) = \frac{|1 - \mu/x|}{b} + \log(2b) \tag{3}$$

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 [20], 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 [20, 44], 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 [10, 20] 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:

$$L_{\text{dropout}}(\theta, p_{drop}) = \frac{1 - p_{drop}}{2N} ||\theta||^2 \quad . \tag{4}$$

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 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.

KITTI Dataset [12]	Type	ALP [%]			ALE [m]		
		$< 0.5m$	$< 1m$	$< 2m$	*Easy*	*Moderate*	*Hard*
Mono3D [4]	Mono	12.6	21.7	35.9	2.10 (2.11)	2.78 (2.96)	3.24 (3.67)
MonoDepth [13]	Mono	20.3	34.8	49.8	1.48 (1.69)	2.23 (3.00)	2.30 (3.48)
Our Geometric baseline	Mono	17.3	32.5	60.9	1.41 (1.47)	1.35 (1.68)	1.54 (1.90)
Our MonoLoco - trained on KITTI	Mono	27.6	48.7	70.3	0.93 (1.03)	1.03 (1.45)	**1.10** (1.61)
Our MonoLoco - trained on nuScenes	Mono	30.4	**51.9**	**71.7**	0.84 (0.92)	**0.91** (1.23)	1.25 (1.74)
3DOP [5]	Stereo	**38.9**	50.0	56.5	**0.59** (0.60)	1.04 (1.03)	1.79 (1.32)

Table 1. Comparing our proposed method against baseline results on KITTI dataset. We calculated ALE for pedestrians commonly detected by all methods to make fair comparison. On parenthesis, we reported the ALE for all the pedestrians detected by each method. We outperform all monocular methods and we achieve comparable performances against 3DOP which leverages on stereo images for training and testing. Our main method uses monocular images and it has not been trained on KITTI. We use PifPaf [23] for 2D poses.

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:

$$Var(\tilde{X}) = \frac{1}{TI}\sum_{t=1}^{T}\sum_{i=1}^{I}\tilde{x}_{t,i}^2(\mu_t, b_t) - \left[\frac{1}{TI}\sum_{t=1}^{T}\sum_{i=1}^{I}\tilde{x}_{t,i}(\mu_t, b_t)\right]^2 \quad (5)$$

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 our model on the *nuScenes* dataset [34]. *NuScenes* is a dataset recently released which contains 24k instances of pedestrians over 6k images sampled at 2 Hz. Even if up to now only a teaser of 100 scenes have been released, it is currently the largest dataset available for 3D object detection. No previous method for monocular pedestrian localization has been evaluated on nuScenes to the best of our knowledge.

The reference dataset for 3D object detection is KITTI [12]. It contains 7481 training images with up to 30 pedestrians per image along with the camera calibration files. We use this dataset for evaluation purposes.

Training Procedure. We split nuScenes into training and test set by scenes and we choose the best model by cross-validation. We run the training procedure for 200 epochs using the Adam optimizer [21], a starting learning rate of 0.005 with exponential decay and mini-batches of 256. For the pose detections we directly use the pre-trained weights of PifPaf or Mask R-CNN trained on COCO [27] and we don't update the weights during training.

We do not apply data augmentation but we modify the resolution of the images to match the minimum dimension of 32 pixels of COCO instances. NuScenes contains high-definition images of 1600x900 pixels and Mask R-CNN resizes each input image to an ideal size of 1333x800 pixels. In order to avoid down-sampling of the images in Mask R-CNN, we first remove the upper side of each image (approximately corresponding to the sky) and then we split each image into two overlapping crops of 1000x600 pixels. We finally apply non-max suppression to avoid double detections. PifPaf does not resize images by default and we double the resolution of the original images. We apply the same procedure for evaluation on KITTI.

The code is developed using PyTorch [37]. Working with a low-dimensional latent representation is very appealing as it allows us to make 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*. [45] for the *car* category. ALP considers a prediction correct if the error between the predicted distance and the ground truth is smaller than a certain threshold. We also analyze the average localization error (ALE) in two different conditions. Following KITTI guidelines, we split the detected pedestrians in three difficulty regimes based on bounding box height, levels of occlusion and truncation: *easy*, *medium* and *hard*. We also analyze the ALE as a function of the distance and we compare the results against the task error of eq 1. The task error defines the target error for monocular approaches due to the ambiguity of the task.

Evaluation Protocol. We train our model on nuScenes and we evaluate it on KITTI dataset. For evaluation, we follow the train/val split of Chen *et al*. [4] and we ignore the training images, only performing evaluation over the 3769 validation images. We train and evaluate our main model on

two different datasets to analyze generalization capabilities of our network. However, for completeness we also show results on a model trained exclusively on KITTI. We do not perform cross-dataset training.

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 3 segments: head to shoulder, shoulder to hip and hip to ankle.

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 equation 2. We calculate the mean and the standard deviation in meters of each of the segments for all the instances in the training set of nuScenes. 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.

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

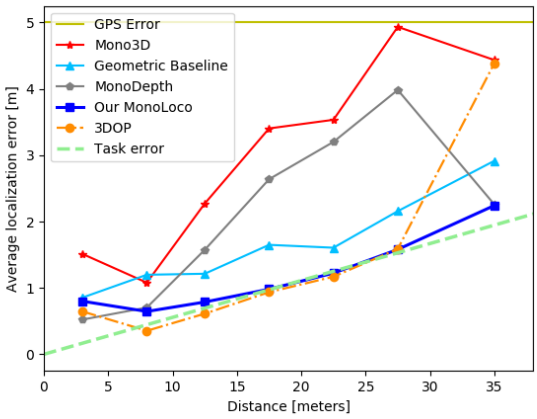

Figure 4. Average localization error for the instances commonly detected by all methods. We outperform the monocular Mono3D [4] while achieving comparable performances with the stereo 3DOP [5]. Monocular performances are bounded by our modeled *task error* in equation 1.

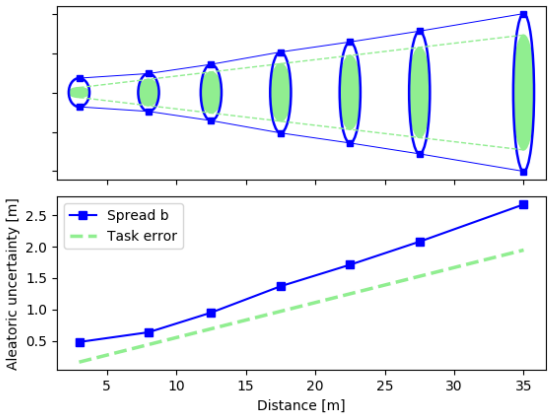

Figure 5. Results of aleatoric uncertainty predicted by MonoLoco (spread b) and the modeled aleatoric uncertainty due to human height variation (task error e). The term $b - e$ is indicative of the aleatoric uncertainty due to noisy observations. On the top figure, we visualize the average predicted and ground truth confidence intervals $\pm b$ and $\pm e$ at various distances.

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

- **Mono3D** [4] 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** [13] 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 close to 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** [5] is a stereo approach for pedestrians, cars and cyclists and their 3D detections are publicly available.

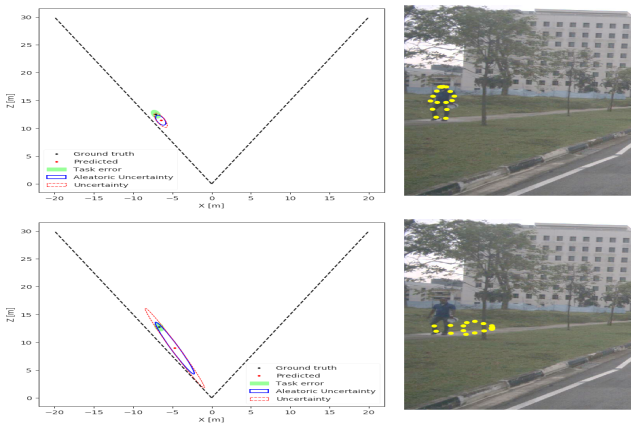

Figure 6. Simulating the outlier case of a person lying on the ground. In the top image, the predicted confidence interval is small and the detection accurate. In the bottom image, we create an outlier pose projecting on the ground the original pose. The network predicts higher uncertainty, a useful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

Mask R-CNN [15]	ALE [m]				
	$\frac{10}{0}$	$\frac{20}{10}$	$\frac{30}{20}$	$\frac{+}{30}$	*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 [23]	ALE [m]				
	$\frac{10}{0}$	$\frac{20}{10}$	$\frac{30}{20}$	$\frac{+}{30}$	*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 2. Impact of different loss functions and pose detectors on nuScenes [34] validation set.

5.3. Results.

Localization Accuracy. Table 1 summarizes our quantitative results on KITTI. We strongly outperform all the other monocular approaches on all metrics. We obtain comparable results with the stereo approach 3DOP [5], which has been trained and evaluated on KITTI and makes use of stereo images during training and test time. On the other hand, our method has been only trained on nuScenes, making it less likely to overfit on KITTI.

In Figure 4, we make an in-depth comparison against monocular Mono3D and the stereo approach 3DOP, analyzing the average localization error as a function of the ground truth distance. We also compare the performances against the *task error* due to human height variations modeled in equation 1, which approximate the upper bound of performances on pedestrian localization for monocular methods. It is worth noticing that 3DOP is able to overcome the threshold for closer instances. This is expected as stereo approaches are not subjected to the same intrinsic limitation of the monocular ones. Even with such intrinsic ambiguity, our method results in more stable performances, with a quasi-linear behaviour which almost replicates the target threshold. Figure 7 shows our qualitative results on challenging images from nuScenes and KITTI datasets.

Uncertainty. We compare in Figure 5 the aleatoric uncertainty predicted by our network through spread b with e, the *task error* due to human height variation defined in equation 1. The predicted spread b is a property of each set of inputs and, differently from 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 [20], i.e. the 2D joints inferred by the pose detector. Hence, we can approximately define the predictive aleatoric uncertainty due to noisy joints as $b - e$ and we observe that the further a person is from the camera, the higher is the term $b - e$. The spread b is the result a probabilistic interpretation of the model. On nuScenes validation set, we observe that 69% of the detected instances lie in the interval $\mu \pm b$. On KITTI validation set, we observe that 68% of the annotations lie in the confidence interval given by the spread b.

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 4. We test different dropout probabilities and choose $p_{\text{drop}} = 0.2$. We perform 100 computationally expensive forward passes and, for each of them, 100 computationally cheap samples from Laplace distribution using eq 5. As a result, 85% and 82% of pedestrians lie inside the predicted confidence intervals for the validation sets of nuScenes and KITTI, respectively.

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 uncertainties 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

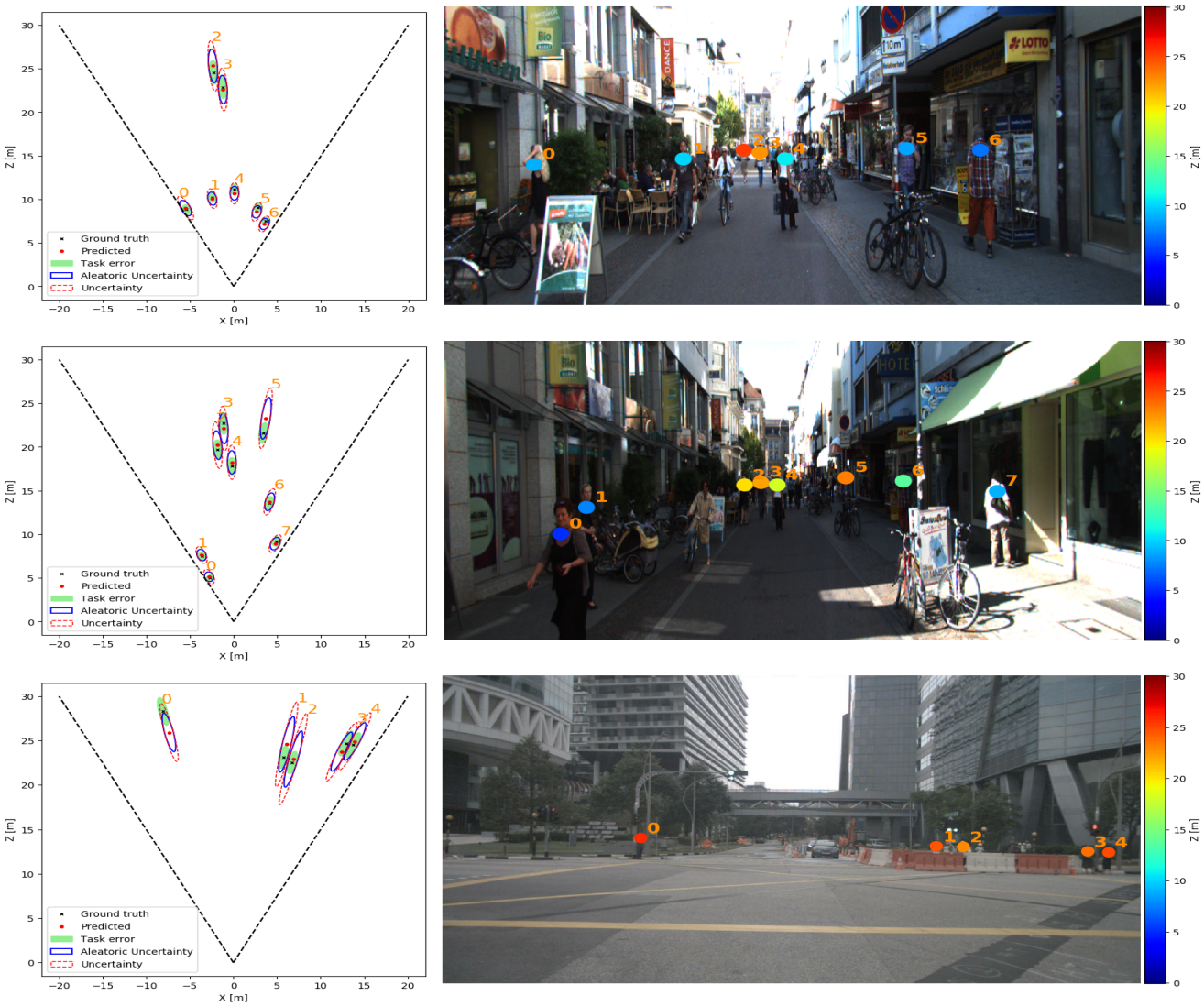

Figure 7. Illustration of results from KITTI [12] (top and middle) and nuScenes [34] (bottom) datasets containing true and inferred distance information as well as confidence intervals (represented by ellipses with minor axis of 1 meter). We observe that the predicted uncertainty increases in case of occlusions (bottom image, pedestrians 1 and 2).

"standard" detection of the same person standing up. When the pedestrian is lying down, the network predicts an unusually large confidence interval which still includes the ground truth location.

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 higher confidence intervals when compared to visible pedestrians at similar distances.

5.4. Ablation studies

In Table 2 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.

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 3D localization.

References

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. 2

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017. 2

[3] F. Chabot, M. A. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1827–1836, 2017. 2

[4] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016. 1, 2, 5, 6

[5] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015. 5, 6, 7

[6] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. 1, 4

[7] H. Fang, S. Xie, and C. Lu. Rmpe: Regional multi-person pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017. 2

[8] D. Feng, L. Rosenbaum, and K. Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3266–3273. IEEE, 2018. 2

[9] J. Freeman, T. Cole, S. Chinn, P. Jones, E. White, and M. Preece. Cross sectional stature and weight reference curves for the uk, 1990. *Archives of disease in childhood*, 73(1):17–24, 1995. 3

[10] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 4

[11] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017. 2

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5, 8

[13] C. Godard, O. Mac Aodha, and G. 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. 5, 6

[14] A. Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011. 2

[15] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2, 3, 7

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Krahenbuhl, T. Darrell, and F. Yu. Joint monocular 3d vehicle detection and tracking. *CoRR*, abs/1811.10742, 2018. 2

[18] S. Huang, M. Gong, and D. Tao. A coarse-fine network for keypoint localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3047–3056, 2017. 2

[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[20] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 1, 2, 4, 7

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018. 2

[23] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. *CVPR*, 2019. 3, 5, 7

[24] M. E. Kundegorski and T. P. Breckon. A photogrammetric approach for real-time 3 d localization and tracking of pedestrians in monocular infrared imagery. 2014. 2, 3

[25] A. Kundu, Y. Li, and J. M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3559–3568, 2018. 2

[26] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017. 2

[27] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 5

[28] F. Manhardt, W. Kehl, and A. Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. *arXiv preprint arXiv:1812.02781*, 2018. 2

[29] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668. IEEE, 2017. 2, 4

[30] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2

[31] J. Mukhoti and Y. Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2

[32] R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, Toronto, Ont., Canada, Canada, 1995. AAINN02676. 2

[33] A. Newell and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 2

[34] NuTonomy. NuScenes data set. https://www.nuscenes.org/, 2018. 5, 7, 8

[35] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018. 2

[36] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy. Towards accurate multi-person pose estimation in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3711–3719, 2017. 2

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 5

[38] M. D. Richard and R. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991. 2

[39] T. Roddick, A. Kendall, and R. Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 1, 2

[40] T. Salimans, D. P. Kingma, M. Welling, et al. Markov chain monte carlo and variational inference: Bridging the gap. In *ICML*, volume 37, pages 1218–1226, 2015. 2

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 2, 4

[42] P. M. Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489, 2008. 2

[43] M. G. Wing, A. Eklund, and L. D. Kellogg. Consumer-grade global positioning system (gps) accuracy and reliability. *Journal of forestry*, 103(4):169–173, 2005. 3

[44] S. Wirges, M. Reith-Braun, M. Lauer, and C. Stiller. Capturing object detection uncertainty in multi-layer grid maps. *arXiv preprint arXiv:1901.11284*, 2019. 4

[45] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3d voxel patterns for object category recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1903–1911, 2015. 2, 5

[46] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Subcategory-aware convolutional neural networks for object proposals and detection. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 924–933. IEEE, 2017. 2

[47] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 2

[48] B. Xu and Z. Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018. 2