Perceiving Humans: From Monocular 3D Localization to Social Distancing

Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi

***Abstract*— Perceiving humans in the context of Intelligent Transportation Systems (ITS) often relies on multiple cameras or expensive LiDAR sensors. In this work, we present a new cost-effective vision-based method that perceives humans' locations in 3D and their body orientation from a single image. We address the challenges related to the ill-posed monocular 3D tasks by proposing a neural network architecture that predicts confidence intervals in contrast to point estimates. Our neural network estimates human 3D body locations and their orientation with a measure of uncertainty. Our proposed solution (i) is privacy-safe, (ii) works with any fixed or moving cameras, and (iii) does not rely on ground plane estimation. We demonstrate the performance of our method with respect to three applications: locating humans in 3D, detecting social interactions, and verifying the compliance of recent safety measures due to the COVID-19 outbreak. We show that it is possible to rethink the concept of "social distancing" as a form of social interaction in contrast to a simple location-based rule. We publicly share the source code towards an open science mission.**

***Index Terms*— Autonomous systems, computer vision, intelligent robots, autonomous vehicles, object detection, social distancing, COVID-19.**

I. Introduction

OVER the past decades, we have witnessed new emerging technologies to localize humans in 3D, ranging from vision-based [1]–[5], to LiDAR-based solutions [6], [7] and multi-sensor approaches [8], [9]. On one hand, vision-based technologies can capture detailed body poses and texture properties, but rely on a costly calibrated network of cameras [10]–[12]. On the other hand, LiDAR sensors are limited by high cost, noise in case of adverse weather, and sparsity of point clouds over long ranges [4], [13], [14]. In this work, we show that given a single cost-effective RGB camera, we can not only extract humans' 3D locations but also their body orientations. Consequently, we can go beyond monocular 3D localization of humans and detect social interactions (*e.g.*, whether two people are talking to each other) in transportation hubs, and even verify compliance with the recent safety measures due to the COVID-19 outbreak.

The COVID-19 pandemic has forced authorities to limit non-essential movements of people, especially in crowded areas or public transport [15]. Social distancing measures are becoming essential to restart passenger services, *e.g.*, leaving train seats unoccupied. Yet in many contexts it is not obvious how to preserve inter-personal distances. When the risk of contagion remains, we should work to minimize it, and perceiving social interactions can play a vital role. In fact, talking with a person does not incur the same risk of infection as passing someone in the street. The infection rate of a disease can be summarized as the product of exposure time and exposure to virus particles [16], [17]. When people are talking together, not only does the exposure time escalate, but the act of speaking itself increases the release of respiratory droplets about tenfold [18], [19]. These analyses urge us to rethink safety measures and focus on proximal social interactions, which can be defined as any behavior of two or more people mutually oriented towards each other and who influence or take into account each other's subjective experiences or intentions [20]. We show that we can monitor the concept of "social distancing" as a form of social interaction in contrast to a simple location-based rule or smartphone-based solutions [21]–[23]. A few methods have studied interactions from images [24], [25], but their results are either limited to personal photos, [26], indoor scenarios, [27], or necessitate a homography calibration [24], [25]. However, the study of social distancing requires an understanding of social interactions in a variety of unconstrained scenarios, either outdoors or within large facilities.

In this paper, we propose a deep learning approach that perceives humans and their social interactions in the 3D space from visual cues only. We argue that the fundamental challenge behind recognizing social interactions from a monocular camera is to perceive humans in 3D, an intrinsically ill-posed problem. We address this ambiguity by predicting confidence intervals in contrast to point estimates through a loss function based on the Laplace distribution. Our approach consists of three main steps. First, we use an off-the-shelf pose detector [28] to obtain 2D keypoints, a low-dimensional representation of humans. Second, the 2D poses are fed into a light-weight feed-forward neural network that predicts 3D locations, orientations and corresponding confidence intervals for each person. Finally, driven by these perception tasks, we aim at investigating how people use the space when interacting in groups. According to the subfield of proxemics, people tend to arrange themselves spontaneously in specific configurations called F-formations [29]. The detection of F-formations is critical to infer social relations [24], [25]. Our intuition is that knowing the 3D location and orientation of people in a scene allows the accurate retrieval

Manuscript received August 10, 2020; revised December 22, 2020 and January 22, 2021; accepted March 11, 2021. This work was supported in part by the Swiss National Science Foundation under Grant 2OOO21-L92326 and in part by the Swiss National Science Foundation (SNSF) Spark fund under Grant 190677. The Associate Editor for this article was Z. Duric. *(Corresponding author: Lorenzo Bertoni.)*

The authors are with the Visual Intelligence for Transportation (VITA) Lab, EPFL, 1015 Lausanne, Switzerland (e-mail: lorenzo.bertoni@epfl.ch).

Digital Object Identifier 10.1109/TITS.2021.3069376

Fig. 1. Our method retrieves 3D locations with confidence intervals, body orientations, social interactions and social distancing in the wild from a single RGB image. We leverage 2D human poses as intermediate representations to verify social distancing compliance while preserving privacy.

of F-formations with simple probabilistic rules. Inspired by [24], [25], we exploit our predicted confidence intervals to develop a simple probabilistic approach to detect F-formations and social interactions among humans. Consequently, we show that we can redefine the concept of social distancing to go beyond a simple measure of distance. We provide simple rules to verify safety compliance in indoor/outdoor scenarios based on the interactions among people rather than their relative position alone. Finally, the design of our pipeline encourages privacy-safe implementations by decoupling the image processing step. Our network is trained on and performs inference with anonymous 2D human poses. An example is provided in Figure 1, where 3D location, orientation and interactions among people are analyzed to verify social distancing compliance in a private manner.

Technically, our main contributions are three-fold: (i) we outperform monocular methods for the 3D localization task on the publicly-available KITTI dataset [30] while also estimating meaningful confidence intervals; (ii) we effectively capture social interactions among people on the Collective Activity Dataset [31] without any additional training or homography estimation; (iii) we show that we can redefine the concept of social distancing based on social cues while preserving the privacy of its users. Our code is publicly available.[1]

II. Related Work

In this work, we tackle the high-level task of understanding 3D spatial relations among humans from a single RGB image without ground plane estimation. The core of our pipeline is composed of a sequence of low-level tasks to process the image and extract 3D information, which can be called monocular 3D vision. The more general field of computer vision has experienced a fundamental transition towards data-hungry deep learning methods thanks to their natural ability to process data in raw form [32]. The transition started with 2D tasks, such as object detection [33], [34] and human pose estimation [35], and it expanded to 3D tasks such as 3D object detection [36]–[38], object recognition [39], depth estimation [40], or even forecasting tasks [41]. A crucial factor in this transformation has been the release of massive datasets for 2D [42]–[44] and 3D tasks [30], [45]–[48], especially in the context of autonomous driving. While perception tasks have been monopolized by relatively new deep learning algorithms, the study of social interactions is based on historic discoveries in behavioural science. In this work, we only focus on *proxemics*: the subfield relating human interactions with the use of space [49]. The remainder of this section is organized as follows. First, we review 2D and 3D tasks that compose our perception pipeline, namely human pose estimation, monocular 3D object detection, and uncertainty estimation. Last, we focus on the study of proxemics and its applications for computer vision and transportation research.

A. Monocular 3D Vision

We include three different sub tasks under the "Monocular 3D Vision" umbrella, as they all contribute to perceive humans in the 3D space from single RGB images. We are interested in algorithms that can operate in outdoor and crowded environments, so when applicable, we focus our review on perception techniques for autonomous driving.

1) 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 and bottom-up methods. Top-down approaches consist in detecting each instance in the image first and then estimating body joints within the boundaries of the inferred bounding box [50]–[53]. Bottom-up approaches estimate separately each body joint through convolutional architectures and then combine them to obtain a full human pose [35], [35], [54]–[56]. More recently PifPaf [28], [57] proposed a method tailored for autonomous driving scenarios that performs well in low-resolution, crowded and occluded scenes. Related to our work is Simple Baseline [58], which shows the effectiveness of latent information contained in 2D joints stimuli. They achieve state-of-the-art results by simply predicting 3D joints from 2D poses through a light, fully connected network. However, these lines of work estimate relative 3D joint positions [59]–[61],

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

or relative 3D meshes [62], [63], not providing any information about the real 3D location in the scene.

2) Monocular 3D Object Detection: The majority of approaches for monocular 3D object detection in the transportation domain focus on vehicles as they are rigid objects with known shape. Very recently, a few works have extended their approaches to the pedestrian category. MonoPSR [64] evaluates pedestrians from monocular RGB images, leveraging point clouds at training time to learn local shapes of objects. MonoDIS [65] proposes to disentangle the contribution of each loss component, while SMOKE [66] combines a single keypoint estimate with regressed 3D variables. Kundegorski and Breckon [67] achieve reasonable performances combining infrared imagery and real-time photogrammetry. Alahi *et al*. combine monocular images with wireless signals [68] or with additional visual priors [10], [69], [70]. The seminal work of Mono3D [36] exploits deep learning to create 3D object proposals for *car*, *pedestrian* and *cyclist* categories but it does not evaluate 3D localization of pedestrians. It assumes a fixed ground plane orthogonal to the camera and the proposals are then scored based on scene priors, such as shape, semantic and instance segmentations. The following methods continue to leverage Convolutional Neural Networks and focus only on *Car* instances. To regress 3D pose parameters from 2D detections, Deep3DBox [71], MonoGRnet [72], and Hu *et al*. [73] use geometrical reasoning for 3D localization, while Multi-fusion [74] and ROI-10D [75] incorporate a module for depth estimation. Roddick *et al*. [76] escape 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 [77]–[80]. While many of the related methods achieve reasonable performances for vehicles, current literature lacks monocular methods addressing other categories in the context of autonomous driving, such as pedestrians and cyclists.

3) Uncertainty Estimation in Computer Vision: Deep neural networks need 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 [81], [82] are 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 [83]–[85] and Deep Ensembles [86]. Alternatively, Gal *et al*. [87], [88] show that applying dropout [89] at inference time yields a form of variational inference where a mixture of multivariate Gaussian distributions with small variances models the network parameters. This technique, called Monte Carlo (MC) dropout, has earned great popularity due to its adaptability to non-probabilistic deep learning frameworks. Very recently, Postels *et al*. [90] proposed a sampling-free method to approximate epistemic uncertainty, treating noise injected in a neural network as errors on the activation values. In computer vision, uncertainty estimation using MC dropout has been applied for depth regression tasks [90], [91], scene segmentation [91], [92] and, more recently, LiDAR 3D object detection for cars [93]. In this work, we demonstrate its relevance for monocular human 3D localization.

B. Social Interactions

We aim to capture social interactions among people and monitor social distancing from visual cues. Related works include the broad field of behavioral science [94]. Here we focus on the subfield called *proxemics*, which investigates how people use and organize the space they share with others [25], [49]. People tend to arrange themselves spontaneously in specific configurations called F-formations [29]. These formations are characterized by an internal empty zone (o-space) surrounded by a concentric ring where people are located (p-space). According to Kendon [29]: "*an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access*".

These formations characterize how people use the space when interacting with each other. They are characterized by three types of social spaces [24], [49]:

1) *o-space:* A circular empty region to preserve the personal space of the participants around it. Every participant looks inward and no people are allowed inside. The type of relation (*e.g*., personal or business-related) defines the dimensions of this space,
2) *p-space:* a concentric ring around the o-space that contains all the participants,
3) *r-space:* the area outside the p-space.

In the case of two participants, typical F-formations are vis-a-vis, L-shape, and side-by-side. For larger groups, a circular formation is typically formed [95]. An example of an F-formation configuration is shown in Figure 3.

To the best of our knowledge, Cristani *et al*. 2011a [24] is the first work to focus solely on visual cues to discover F-formations and social interactions. In parallel, Cristani *et al*. 2011b [25] study how people get closer to each other when the social relation is more intimate. The following works have proposed various techniques to automatically detect F-formations in heterogeneous real crowded scenarios [96]–[99]. In all approaches, it is clear how the detection of F-formations is critical to infer social relations and we decide to follow their lead. This line of work, however, considers as input the position of people on the ground floor and their orientation [25] or requires a homography estimation to compute the x-y-z coordinates of humans [24]. On the contrary, our approach works end-to-end from a single RGB image. The perception stage, *i.e*., extracting 3D detections from a monocular image, is arguably the most challenging one due to the intrinsic ambiguity of perspective projections.

Finally, social interactions have also been studied in the context of personal photos [26] or egocentric photo-streams [27], [100], [101]. Both approaches assume humans to stand less than a few meters apart from each other and the camera, and do not scale to long range applications, such as monitoring an airport terminal. Recently, deep learning approaches have been adopted to understand social interactions under a different perspective. Joo *et al*. [94] learn to predict behavioral cues

of a target person (*e.g.*, body orientation) from the position and orientation of another person. They learn the dynamics between social interactions in a data-driven manner, laying the foundations for deep learning to be applied in the field of behavioral science.

III. 3D Localization Ambiguity

A critical challenge in understanding social interactions from visual cues is the 3D localization pillar. Inferring distance of humans from monocular images is a fundamentally ill-posed problem. The majority of previous works have circumvented this challenge by assuming a planar ground plane and estimating a homography by manual measurement or by knowing some reference elements [24], [36], [103], [104]. These approaches do not work when people are on stairs and they require a static calibrated setup. In this work, we address their limitations by directly estimating distance of humans without relying on a ground plane or homography. This problem is ill-posed due to human variation of height. If every pedestrian has the same height, there would be no ambiguity. However, does this ambiguity prevent from robust localization? This section is dedicated to explore this question and analyze the maximum accuracy expected from monocular pedestrian localization.

We are interested in the 3D localization error due to the ambiguity of perspective projection. Our approach consists in assuming that all humans have the same height h_{mean} and analyzing the error of this assumption. Inspired by Kundegorski and Breckon [67], we model the localization error related to height variation 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 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 178 *cm* for males and 165 *cm* for females with a standard deviation of around 7 *cm* in both cases [105]. 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 [106], 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

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

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% [67], [106]. 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 [67]. Figure 4 shows the expected localization error $\hat{e}$ due to height variations in different cases as a linear function of the ground-truth distance from the camera d_{gt}. For a pedestrian 20 meters far, the localization error is approximately 1 meter. This analysis shows that the ill-posed problem of localizing humans, while imposing an intrinsic limit, does not prevent from a good enough localization in many applications.

IV. Proposed Method

The goals of our method are (i) to detect humans in 3D given a single image and (ii) to leverage this information to recognize social interactions and monitor social distancing. Figure 2 illustrates our overall method, which consists of three main steps. First, we exploit a pose detector to escape the image domain and reduce the input dimensionality. 2D human 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 that predicts x-y-z coordinates and the associated uncertainty, orientation, and dimensions of each pedestrian. In the training phase, there is no supervision for the localization ambiguity. The network implicitly learns it from the data distribution. Third, the network estimates are combined to obtain F-formations [49] and recognize social interactions.

A. 3D Human Detection

The task of 3D object detection is defined as detecting 3D location of objects along with their orientation and dimensions [30], [45]. The ambiguity of the task derives from the localization component as described in Section III. Hence, 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. 2 allows us to compare the predicted confidence intervals with the intrinsic ambiguity of the task.

1) 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 $\left[x_i^*, y_i^*, 1\right]^T$ using the camera intrinsic matrix K:

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

This transformation is essential to prevent the method from overfitting to a specific camera.

2) 2D Human Poses: We obtain 2D joint locations of humans using the off-the-shelf pose detector PifPaf [28], [57], a state-of-the-art, bottom-up method designed for crowded scenes and occlusions. The detector can be regarded as a stand-alone module independent from our network, which uses 2D joints as inputs. PifPaf has not been fine-tuned on any additional dataset for 3D object detection as no annotations for 2D poses are available.

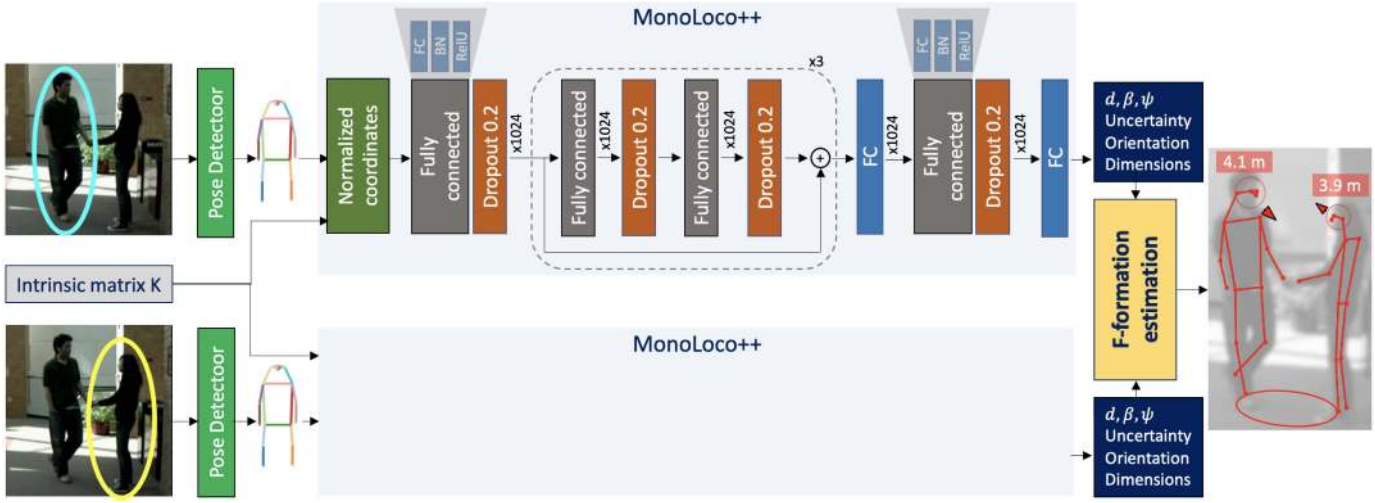

Fig. 2. Overall architecture. **MonoLoco++**: the input is a set of 2D joints extracted from a raw image and the output is the 3D location, orientation and dimensions of a human with the localization uncertainty. 3D location is estimated with spherical coordinates: radial distance d, azimuthal angle β, and polar angle ψ. Every fully connected layer (FC) outputs 1024 features and is followed by a Batch Normalization layer (BN) [102] and a ReLU activation function. **F-formations**: all estimates from MonoLoco++ are analyzed with an *all-vs-all* approach to discover F-formations and estimate social interactions/distancing using Eq. 8.

3) Output: We predict 3D localization, dimensions, and viewpoint angle with a regressive model. Estimating depth is arguably the most critical component in vision-based 3D object detection due to intrinsic limitations of monocular settings described in Section III. However, due to perspective projections, an error in depth estimation z would also affect the horizontal and vertical components x and y. To disentangle the depth ambiguity from the other components, we use a spherical coordinate system (d, β, ψ), namely radial distance d, azimuthal angle β, and polar angle ψ. Another advantage of using a spherical coordinate system is that the size of an object projected onto the image plane directly depends on its radial distance d and not on its depth z [5]. 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.

As already noted in [107], the viewpoint angle is not equal to the object orientation as people at different locations may share the same orientation θ but results in different projections. Hence, we predict the viewpoint angle α, which is defined as $\alpha = \theta + \beta$, where β denotes the azimuth of the pedestrian with respect to the camera. Similarly to [107], we also parameterize the angle as $[\sin\alpha, \cos\alpha]$ to avoid discontinuity. Regarding bounding box dimensions, we follow the standard procedure to calculate width, height and length of each pedestrian. We calculate average dimensions from the training set and regress the displacement from the expectation.

4) Minimization Objective: Our final loss is the logarithm of the probability that all components are "well" predicted, *i.e.*, it is the sum of the log-probabilities for the individual components. For every component but the 3D localization, we use a vanilla L_1 loss. To regress distances of people, we use a Laplace-based L1 loss [91], which we describe in the following section. Our minimization objective is a simple non-weighted sum of each loss function.

5) Base Network: The building blocks of our model are shown in Figure 2. The architecture, inspired by Martinez *et al.* [58], is a simple, deep, fully-connected network with six linear layers with 1024 output features. It includes dropout [89] after every fully connected layer, batch-normalization [102] and residual connections [108]. The model contains approximately 8M training parameters.

6) MonoLoco++ vs MonoLoco: We refer to our method as MonoLoco++. Technically, it differs from the previous MonoLoco [5] by:

- the multi-task approach that combines 3D localization, orientation and bounding-box dimensions,
- the use of spherical coordinates to disentangle the ambiguity in the 3D localization task,
- an improved neural network architecture.

Combining precise 3D localization and orientation paves the way for activity recognition and social distancing, which was not possible using MonoLoco [5]. As illustrated in Fig 2, multiple MonoLoco++ estimates are combined into the *F-formation estimation* block to detect social interactions and social distancing. In addition, we will show how the above technical improvements benefit the monocular 3D localization task itself.

B. Uncertainty

In this work, we propose a probabilistic network which models two types of uncertainty: *aleatoric* and *epistemic* [91], [109]. 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.

1) 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:

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

where x represents the ground-truth distance, d the predicted distance, and b the spread, making this training objective an attenuated L_1-type loss via spread b.

During training, the model has the freedom to ignore noisy data and attenuate its gradients by predicting a large spread b. As a consequence, inputs with high uncertainty have a small effect on the loss, making the network more robust to noisy data. The uncertainty is estimated in an unsupervised way, since no supervision is provided. At inference time, the model predicts a Laplace distribution parameterized by the distance d and a spread b. The latter one indicates the model's confidence about the predicted distance. Following [91], 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 [91], [110], we design a Laplace loss that works with relative distances to take into account the role of distance in our predictions. For example in autonomous driving scenarios, 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.

2) Epistemic Uncertainty: To model epistemic uncertainty, we follow [87], [91] 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:

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

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 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 d and spread b from multiple forward passes with MC dropout:

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

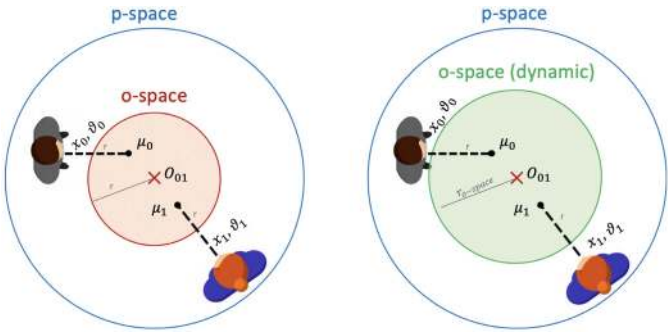

Fig. 3. Illustration of the o-space discovery using [24] on the left and our approach on the right. Both approaches use the candidate radius r to find the center of the o-space, as infinite number of circles could be drawn from two points. Differently from [24], once a center is found, we dynamically adapt the final radius of the o-space $r_{o-space}$ depending on the effective location of the two people.

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

C. Social Interactions and Distancing

We identify social interactions by recognizing the spatial structures that define F-formations (see Section II-B for more details). Our approach considers groups of two people in an "all-vs-all" fashion by studying all the possible pairs of people in an image.

Ideally, two people talking to each other define the same o-space by looking at its center. In practice, 3D localization and orientation of people are noisy and previous methods [24], [25] have adopted a voting approach. They define a candidate radius r of the o-space and each person vote for a center. The average result defines the center of the o-space. In Cristani *et al*. [24], the candidate radius r remains the final radius of the o-space and is fixed for every group of people. However, once the o-space center is found, nothing prevents us from considering its radius $r_{o-space}$ dynamically as the minimum distance between the center and one of the two people. An illustration of the differences is show in Figure 3. Therefore, given the location of two people in the x-z plane $\mathbf{x}$ and their body orientation θ, we define the center and the radius of the o-space as:

$$\begin{aligned} \mathbf{O_{01}} &= \frac{\boldsymbol{\mu_0} + \boldsymbol{\mu_1}}{2} \\ r_{o-space} &= min(|\mathbf{O_{01}} - \mathbf{x_0}|, |\mathbf{O_{01}} - \mathbf{x_1}|) \quad , \end{aligned} \tag{7}$$

where $\mathbf{O_{01}}$ and $r_{o-space}$ are the center and radius of the resulting o-space, $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$ indicate the location of the two candidate centers of the o-space. In general, $\boldsymbol{\mu} = [x + r * cos(\theta), z + r * sin(\theta)]$ and is parametrized by the candidate radius r, which depends on the type of relation (intimate, personal, business, etc.) [49].

Once the o-space is drawn, we verify the following conditions:

$$\begin{aligned} &(a) \quad |\mathbf{x_0} - \mathbf{x_1}| < D_{max} \\ &(b) \quad |\mathbf{O_{01}} - \mathbf{x_i}| < r_{o-space} \ \ \forall i \neq 0, 1 \\ &(c) \quad |\boldsymbol{\mu_0} - \boldsymbol{\mu_1}| < R_{max} \end{aligned} \tag{8}$$

where D_{max} and R_{max} are the maximum distances between two people, and between the candidate centers of the o-spaces, respectively. Vectors are represented in bold.

The above conditions verify the presence of an F-formation, as:

(a) examines whether two people stand closer than a maximum distance D_{max}, *i.e.*, they lie inside a p-space,
(b) examines whether the o-space is empty (no-intrusion condition),
(c) examines whether the two people are looking inward the o-space.

We note that condition (c) is empirical as looking inward is a generic requirement. Two people usually look at each other when talking, but the needs for social distancing may be different. Our goal is not to find perfect empirical parameters for F-formations discovery, but rather to show the effectiveness of combining simple rules and estimating 3D localization and orientation. We consider two people as interacting with each other if the three conditions are verified. This method is automatically extended to larger groups as two people can already cover any possible F-formation (vis-a-vis, L-shape and side-by-side), while three or more people usually form a circle [24]. Further, we are not interested in defining the components of each group, but rather whether people are interacting or not.

1) Social Distancing: The procedure to monitor social distancing can either follow the same steps, or can be adapted to a different context. Risk of contagion strongly increases if people are involved in a conversation [17], [19]. Therefore, recognizing social interactions lets the system only warn those people that incur the highest risk of contagion. In crowded scenes, this is crucial to prevent an extremely high number of false alarms that could undermine any benefit of the technology. Yet social distancing conditions can also be differentiated from the social interaction ones. For example, a third person invading the o-space could mean that the three people involved are not conversing, but still they may be at risk of contagion due to the proximity. How strict these rules should be can only be decided case by case by the competent authority. Our goal is to help assessing the risk of contagion not only through distance estimation but also by leveraging social cues.

2) Uncertainty for Social Interactions: A deterministic approach can be very sensitive to small errors in 3D localization and orientation, which we know are inevitable due to the perspective projection. Therefore, we introduce a probabilistic approach that leverages our estimated uncertainty to increase robustness towards 3D localization noise. We note that Cristani *et al*. [24] also adopted a probabilistic approach injecting uncertainty in a Hough-voting procedure. However, the chosen parameters were driven by sociological and empirical considerations. In our case, uncertainty estimates come directly as an output of the neural network and they are unique for each person. Recalling that the location of each person is defined as a Laplace distribution parametrized by d and b in Eq. 4, we draw k samples from the distribution. For each pair of samples, we verify the above conditions for social interactions. Combining all the results, we evaluate the final probability for a social interaction to occur.

V. Experiments

To the best of our knowledge, no dataset contains 3D labels as well as social interactions or social distancing information. Hence, we used multiple datasets to evaluate monocular 3D localization, social interactions and social distancing separately. The following sections serve this purpose.

A. Monocular 3D Localization

1) Datasets: We train and evaluate our monocular model on the KITTI Dataset [30]. 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 [45] and we test it on KITTI. We do not perform cross-dataset training.

2) 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 1799 training instances for KITTI and 8189 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. Once the human poses are detected, we apply horizontal flipping to double the instances in the training set.

We follow the KITTI train/val split of Chen *et al*. [36] and we run the training procedure for 200 epochs using the Adam optimizer [111], a learning rate of 10^{-3} and mini-batches of 512. The code, available online, is developed using PyTorch [112]. Working with a low-dimensional representation is very appealing as it allows fast experiments with different architectures and hyperparameters. The entire training procedure requires around two minutes on a single GTX1080Ti GPU.

3) Evaluation Metrics: Following [5], we use two metrics to analyze 3D pedestrian localization. First, we consider a prediction as correct if the error between the predicted distance and the ground-truth is smaller than a threshold. We call this metric Average Localization Accuracy (ALA). We use 0.5 meters, 1 and 2 meters as thresholds. We also analyze the average localization error (ALE). To make fair comparison we set the threshold of the methods to obtain similar recall. Compared to [5], we do not evaluate on the common set of detected instances. Their evaluation is not reproducible as the common set depends on the methods chosen for evaluation. In contrast, analyzing ALE and recall allows for simple but fair comparison. Following KITTI guidelines, we assign to each instance a difficulty regime based on bounding box height, level of occlusion and truncation: *easy*, *moderate* and *hard*. However in practice, each category includes instances from the simpler categories, and, due to the predominant number of easy instances (1240 *easy* pedestrians, 900 *moderate* and 300 *hard* ones), the metric can be misleading and underestimate the impact of challenging instances. Hence, we evaluate each instance as belonging only to one category and add the category *all* to include all the instances.

TABLE I

COMPARING OUR PROPOSED METHOD AGAINST BASELINE RESULTS ON THE KITTI DATASET [30]. WE USE PIFPAF [28] AS OFF-THE-SHELF NETWORK TO EXTRACT 2D POSES. FOR THE ALE METRIC, WE SHOW THE RECALL BETWEEN BRACKETS TO INSURE FAIR COMPARISON. WE SHOW RESULTS BY TRAINING WITH THREE DIFFERENT DATA SPLITS: KITTI DATASET [30], NUSCENES TEASER [45] OR A SUBSET OF NUSCENES TO MATCH THE NUMBER OF INSTANCES OF THE KITTI DATASET. ALL CASES SHARE THE SAME EVALUATION PROTOCOL. THE MODELS TRAINED ON NUSCENES SHOW CROSS-DATASET GENERALIZATION PROPERTIES BY OBTAINING COMPARABLE RESULTS IN THE ALE METRIC

Method	Dataset Training	Number of Instances	ALE (m) ↓ [Recall (%) ↑]				ALA (%) ↑		
			$Easy$	$Mod.$	$Hard$	All	$< 0.5m$	$< 1m$	$< 2m$
Mono3D [36]	KITTI	1799	2.26 [89%]	3.00 [65%]	3.98 [34%]	2.62 [69%]	13.0	22.9	38.2
MonoPSR [64]	KITTI	1799	0.88 [96%]	1.86 [68%]	1.85 [16%]	1.19 [69%]	31.1	44.2	57.4
SMOKE [66]	KITTI	1799	0.75 [59%]	1.30 [30%]	1.53 [10%]	0.91 [39%]	18.7	27.3	34.5
MonoDIS [65]	KITTI	1799	**0.66** [85%]	1.26 [64%]	1.83 [32%]	0.93 [66%]	33.2	47.6	57.6
3DOP [104] (Stereo)	KITTI	1799	0.67 [88%]	1.19 [64%]	1.93 [37%]	0.94 [69%]	40.6	53.7	61.4
Geometric [5]	KITTI	-	1.05 [89%]	0.95 [63%]	1.34 [31%]	1.04 [68%]	23.5	41.9	59.4
MonoLoco [5]	KITTI	1799	0.95 [89%]	0.98 [64%]	1.11 [31%]	0.97 [68%]	25.3	43.4	60.5
MonoLoco [5]	nuScenes	8189	0.91 [92%]	1.16 [80%]	1.45 [30%]	1.08 [74%]	27.6	46.6	63.7
Our MonoLoco++	KITTI	1799	0.69 [90%]	**0.71** [66%]	1.37 [31%]	**0.76** [70%]	**37.4**	**53.2**	63.6
Our MonoLoco++	nuScenes	1799	0.81 [92%]	0.84 [68%]	1.14 [29%]	0.84 [70%]	31.8	50.2	63.9
Our MonoLoco++	nuScenes	8189	0.72 [91%]	0.77 [68%]	**1.03** [29%]	**0.76** [70%]	32.5	51.9	**65.6**

4) Geometric 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. 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. We also average between left and right keypoint values to take into account noise in the 2D joint predictions. The result is a single height Δy_{1-2} that 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 [105] described in Section 3 with the anthropometry study of Drillis *et al*. [113], 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.

5) Other Baselines: We compare our monocular method on KITTI against five monocular approaches and a stereo one:

- *MonoLoco*. We compare our approach with MonoLoco [5]. Our MonoLoco++ uses a multi-task approach to learn orientation, has a different architecture and uses spherical coordinates for distance estimation. Both methods share the same off-the-shelf pose detector [28], [57].
- *Mono3D* [36] is a monocular 3D object detector for cars, cyclists and pedestrians. 3D localization of pedestrians is not evaluated but detection results are publicly available
- *MonoPSR* [64] is a monocular 3D object detector that leverages point clouds at training time to learn shapes of objects. In contrast, our method does not use any privileged signal at training time.
- *MonoDIS* [65] is a very recent multi-class 3D object detector that provides evaluations for the pedestrian category on the KITTI dataset.
- *SMOKE* [66] is a single-stage monocular 3D object detection method which is based on projecting 3D points onto the image plane. The authors have shared their quantitative evaluation.
- *3DOP* [104] is a stereo approach for pedestrians, cars and cyclists and their 3D detections are publicly available.

Finally, in Figure 4 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.

B. Monocular Results

1) Localization Accuracy: Table I summarizes our quantitative results on KITTI. We strongly outperform all the other monocular approaches on all metrics and obtain comparable results with the stereo approach 3DOP [104], which has been trained and evaluated on KITTI and makes use of stereo images during training and test time. In addition, we show cross-dataset generalization properties by training our network on a subset of the nuScenes dataset containing only 1799 instances and evaluating it on the KITTI dataset.

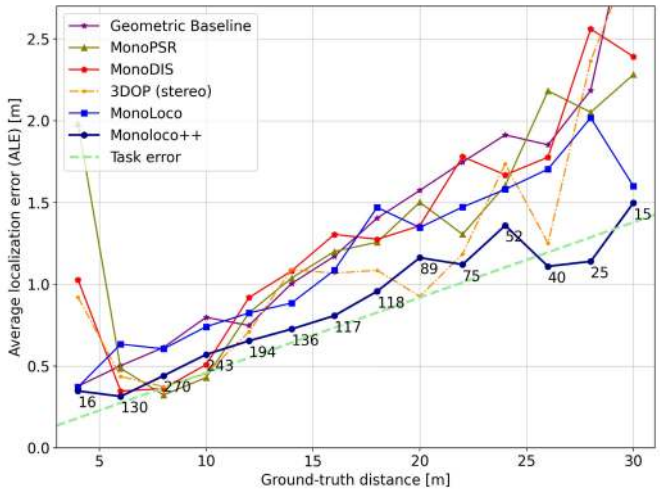

Fig. 4. Average localization error (ALE) as a function of distance. We outperform the monocular MonoPSR [64] and MonoDIS [65], while even achieving more stable results than the stereo 3DOP [104]. Monocular performances are bounded by our modeled task error in Eq. 2. The task error is only a mathematical construction not used in training and yet it strongly resembles the network error, especially for the more statistically significant clusters (number of predicted instances included).

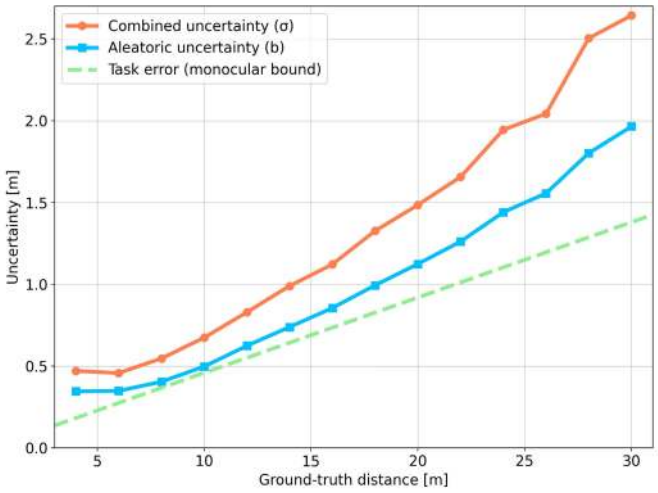

Fig. 5. Aleatoric uncertainties predicted by MonoLoco++ (spread b), and due to human height variation (task error $\hat{e}$) as a function of the ground-truth distance. The term $b - \hat{e}$ is indicative of the aleatoric uncertainty due to noisy observations. The combined uncertainty σ accounts for aleatoric and epistemic uncertainty and is obtained applying MC Dropout [87] at test time with 50 forward passes.

Its generalization properties can be attributed to the low-dimensional input space of 2D keypoints [114].

In Figure 4, we make an in-depth comparison 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 2. Our method results in stable performances that almost replicate the target threshold. More generally, it is notable that the error of each method shows a quasi-linear behaviour. At a short range, the majority of methods show large errors, as the instances are not fully visible in the image. Since our method reasons with keypoints, its performances are more stable. At the 25-30m range MonoLoco error is slightly lower than the task error. This is mainly caused by the statistical fluctuations due to the small sample sizes at those distances. Figure 6 and 7 show qualitative results on challenging images from the KITTI and nuScenes datasets, respectively.

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. While $\hat{e}$ is a linear function of the distance from the camera, the predicted aleatoric uncertainty (through the spread b) is a property of each set of inputs. In fact, b includes not only the uncertainty due to the ambiguity of the task but also the uncertainty due to noisy observations [91], *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 the KITTI validation set, they include 68% of the instances.

3) Combined Uncertainty: The combined aleatoric and epistemic uncertainties are captured by sampling from multiple Laplace distributions using MC dropout. During each of the forward passes, we draw and accumulate samples from the estimated Laplace distribution. Then, we calculate the combined uncertainty as the sample variance of predicted distances in Eq. 6. The magnitude of the uncertainty depends on the chosen dropout probability p_{drop} in Eq. 5. In Table II, 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 a Laplace distribution using Eq. 6. As a result, 84% of pedestrians lie inside the predicted confidence intervals for the validation set of KITTI.

One of our goals is robust 3D estimates for pedestrians, 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. To simulate an autonomous driving scenario, 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.

4) Challenging Cases: We qualitatively analyze the role of the predicted uncertainty in case of an outlier in Figure 9. In the top image, a person is partially occluded and this is reflected in a larger confidence interval. Similarly in the bottom figure, we estimate the 3D localization of a driver inside a truck. The network responds to the unusual position of the 2D joints with a very large confidence interval. In this case the prediction is also reasonably accurate, but in general an unusual uncertainty can be interpreted as a useful indicator to warn about critical samples.

We also show the advantage of estimating distances without relying on homography estimation or assuming a fixed ground plane, such as [36], [104]. The road in Figure 9 (top) is uphill

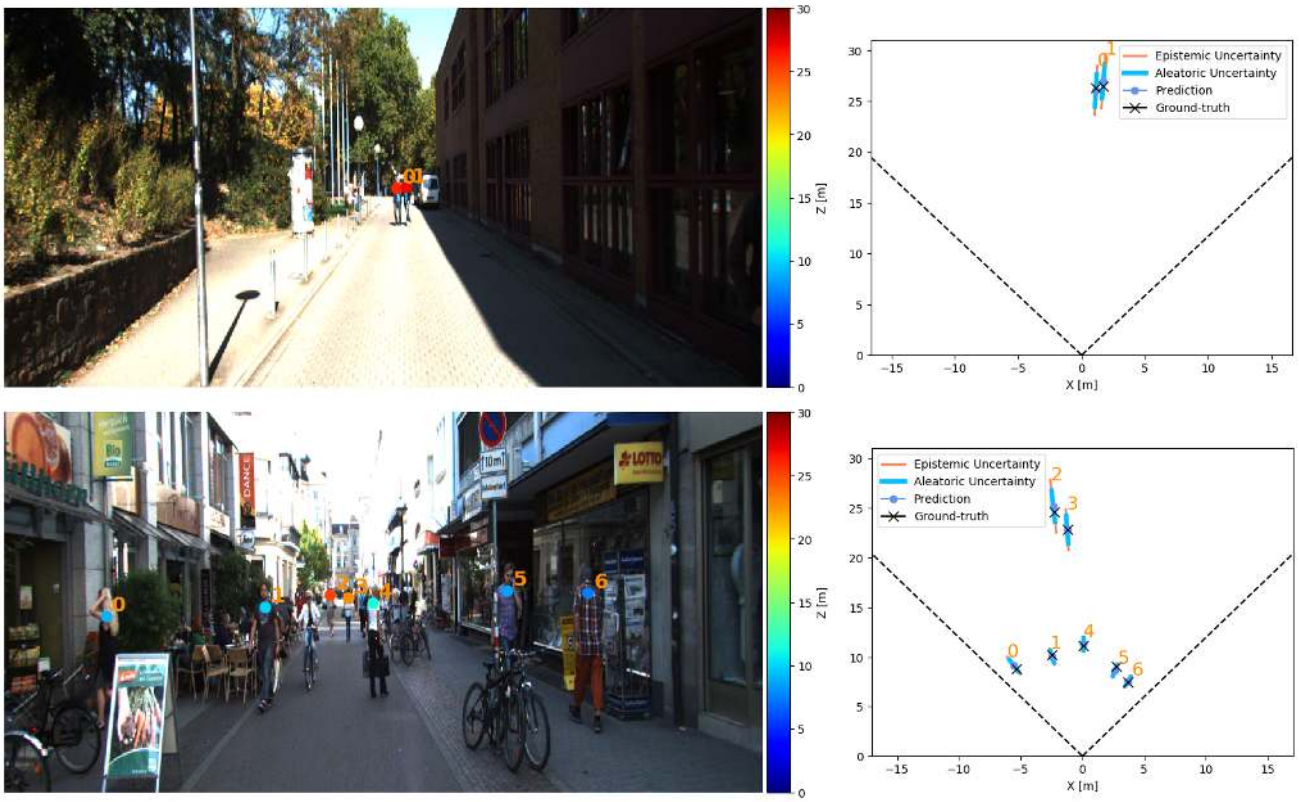

Fig. 6. Qualitative results from the KITTI [30] dataset containing true and inferred distance information as well as confidence intervals. The direction of the line is radial as we use spherical coordinates. Only pedestrians that match a ground-truth are shown for clarity.

TABLE II

PRECISION AND RECALL OF UNCERTAINTY FOR THE KITTI VALIDATION SET WITH 50 STOCHASTIC FORWARD PASSES. $|x-d|$ 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

	$\lvert x-d\rvert/\sigma$	$\lvert\sigma-e\rvert$ [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 III

IMPACT OF DIFFERENT LOSS FUNCTIONS WITH MASK R-CNN [52] AND PIFPAF [28] POSE DETECTORS ON NUSCENES TEASER VALIDATION SET [45]. WE ALSO SHOW RESULTS USING THE AVERAGE LOCALIZATION ERROR (ALE) METRIC AS A FUNCTION OF THE GROUND-TRUTH DISTANCE USING CLUSTERS OF 10 METERS

Mask R-CNN [52]	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 [28]	**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**

as frequently happens in the real world (*e.g.*, San Francisco). MonoLoco++ does not rely on ground plane estimation, making it robust to such cases.

5) Ablation Studies: In Table III, 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. We compare our Laplace-based $L1$ loss of Eq. 4 with a relative L_1 loss

$$L_1(x|d) = |1 - d/x| \quad , \tag{9}$$

and a Gaussian loss

$$L_{\text{Gaussian}}(x|d,\sigma) = \frac{(1-d/x)^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2). \tag{10}$$

The Gaussian Loss is based on the negative log-likelihood of a Gaussian distribution and corresponds to an L_2 loss attenuated by a predicted σ in the location. Intuitively, L_2 type losses are more sensitive to outliers due to their quadratic component. All the losses make use of relative distances for consistency with Eq. 4. From Table III, we observe that L_1-type losses perform slightly better than the Gaussian loss, but the main improvement is given by choosing PifPaf as pose detector.

6) Run Time: A run time comparison is shown in Table IV. Our method is faster or comparable to all the other methods, achieving real-time performance.

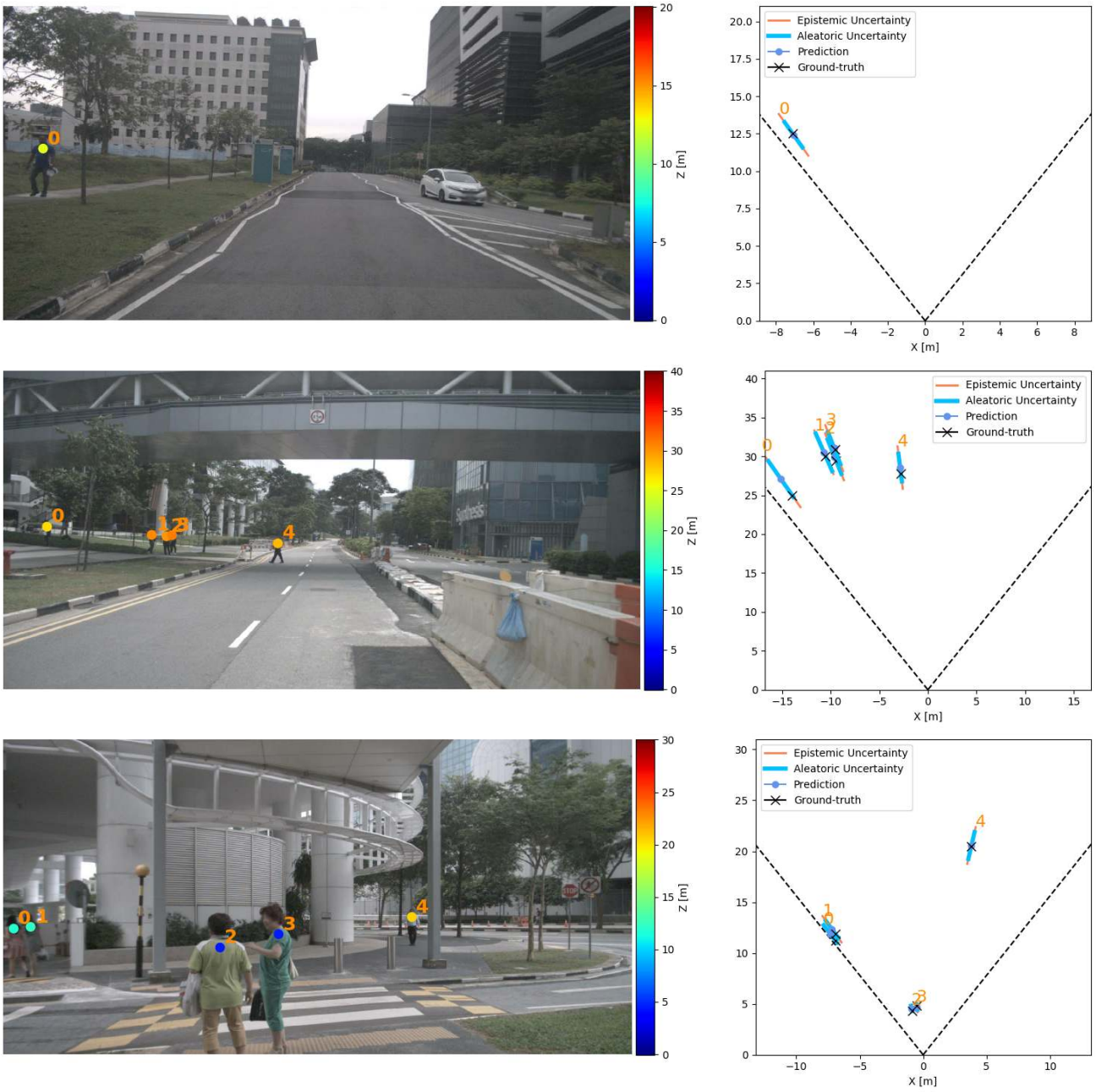

Fig. 7. Qualitative results from the nuScenes dataset [45] containing true and inferred distance information as well as confidence intervals.

C. Social Interactions

To evaluate social interactions we focus on the activity of *talking*, which is considered as the most common form of social interaction [24]. From single images, we evaluate how well we recognize whether people are talking or just passing by, walking away etc.

1) Datasets: We evaluate social interactions on the Collective Activity Dataset [31], which contains 44 video sequences of 5 different collective activities: *crossing*, *walking*, *waiting*, *talking*, and *queuing* and focus on the *talking* activity. The *talking* activity is recorded for both indoor and outdoor scenes, allowing us to test our 3D localization performance on different scenarios. Compared to other deep learning methods [115]–[117], we analyze each frame independently with no temporal information, and we do not perform any training for this task, using all the dataset for testing.

2) Evaluation: For each person in the image, we estimate his/her 3D localization confidence interval and orientation. For every pair of people we apply Eq. 7 and Eq. 8 to discover the F-formation and assess its suitability. We use the following parameters in meters: $D_{max} = 2$ as maximum distance, and $r_1 = 0.3$, $r_2 = 0.5$ and $r_3 = 1$ as radii for o-space candidates. These choices reflect the average distances of *intimate relations*, *casual/personal relations* and *social/consultive* relations, respectively [49].

How much people should look inward the o-space (to assume they are talking) is also an empirical evaluation. We set the maximum distance between two candidate centers

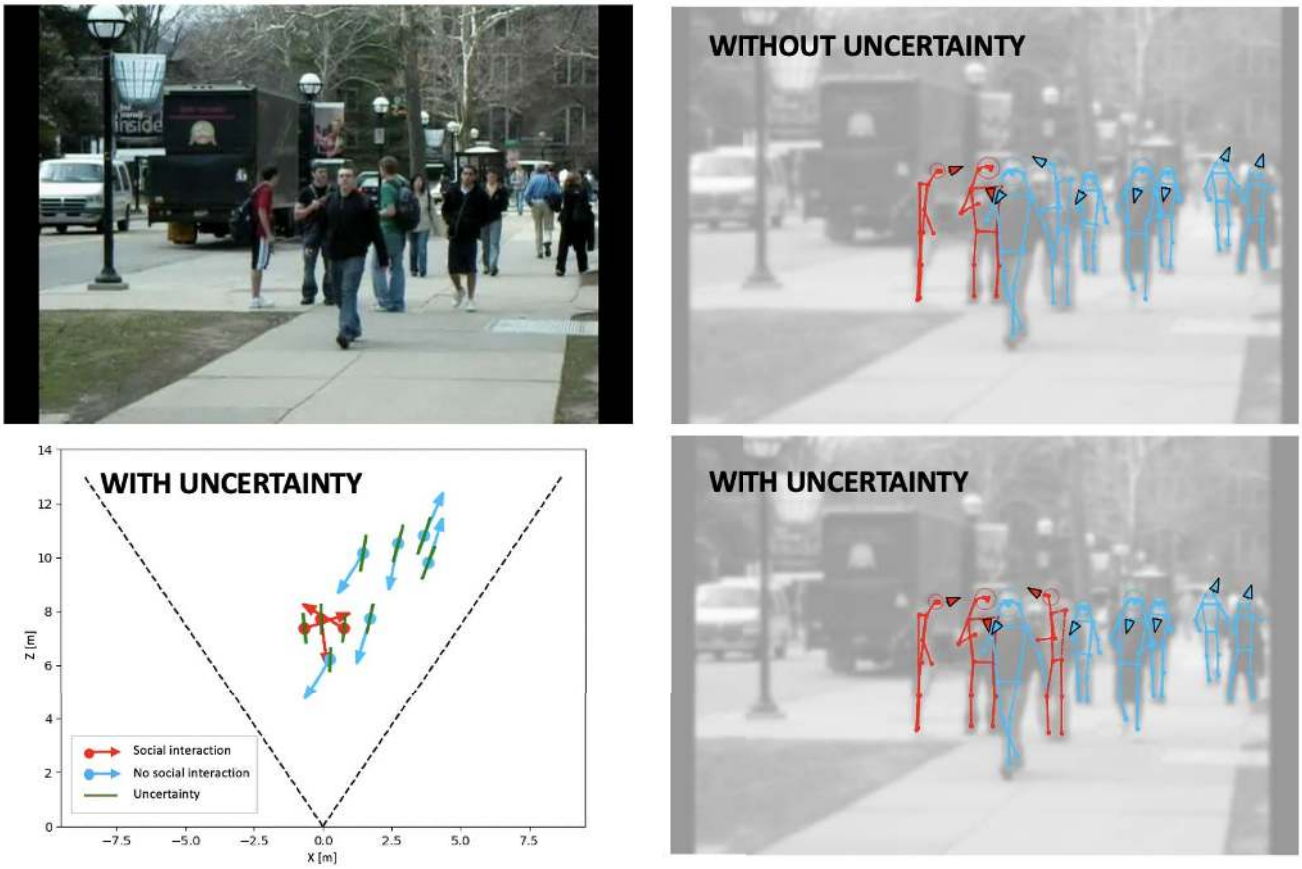

Fig. 8. Estimating whether people are talking to each other (social interaction). The use of uncertainty makes the method more robust to 3D localization errors and improves the accuracy. The bird eye view shows the estimated 3D locations and orientations of all the people. The color of the arrows indicates whether people are involved in talking.

TABLE IV

SINGLE-IMAGE INFERENCE TIME ON A SINGLE GTX 1080Ti FOR THE KITTI DATASET [30] WITH PIFPAF [28] AS POSE DETECTOR. MOST COMPUTATION COMES FROM THE POSE DETECTOR (RESNET 50 / RESNET 152 BACKBONES). FOR THIS STUDY, WE USE ALL THE IMAGES AT THEIR ORIGINAL SCALE THAT CONTAIN AT LEAST A PEDESTRIAN. FOR MONO3D, 3DOP AND MONOPSR WE REPORT PUBLISHED STATISTICS ON A TITAN X GPU. IN THE LAST LINE, WE CALCULATE EPISTEMIC UNCERTAINTY THROUGH 50 SEQUENTIAL FORWARD PASSES. IN FUTURE WORK, THIS COMPUTATION CAN BE PARALLELIZED

Method \ Time [ms]	t^{pose}	t^{model}	t^{total}
Mono3D [36]	-	1800	1800
3DOP [104]	-	2000	2000
MonoPSR [64]	-	200	200
Our MonoLoco++ (1 sample)	89 / 162	10	**99 / 172**
Our MonoLoco++ (50 samples)	89 / 162	51	140 / 213

$R_{max} = r_{o-space}$ for simplicity. We treat the problem as a binary classification task and evaluate the detection recall and the accuracy in estimating whether the detected people are talking to each other. To disentangle the role of the 2D detection task, we report accuracy on the instances that match a ground-truth. To avoid class imbalance, we only analyze sequences that contain at least a person talking in one of their frames. Consequently, we evaluate a total of 4328 instances, of which 52.8% are *talking*.

3) Voting Procedure: To account for noise in 3D localization, we sample our results from the estimated Laplace distribution parameterized by distance d and spread b (Eq. 4). Each sample votes for a candidate center $\boldsymbol{\mu}$ and we accumulate the voting. If an agreement is reached within at least 25% of the samples, we consider the target pair of people as involved in a social interaction and/or at risk of contagion. MonoLoco++ estimates a unique spread b for each pedestrian, which accounts for occlusions or unusual locations, as seen in Figure 9. Further, we compare this technique to (i) a baseline approach that leverages 3D localization but not orientation, (ii) a deterministic approach that does not include uncertainty, and (iii) a probabilistic approach where the uncertainty is provided by the task error defined in Eq. 2.

4) Results: Table V shows the results for the *talking* activity in the Collective Activity Dataset [31]. Our MonoLoco++ detects whether people are talking from a single RGB image with 91.5% accuracy without being trained on this dataset, but only using the estimated 3D localization and orientation. The uncertainty estimation plays a crucial role in dealing with noisy 3D localizations as shown in the ablation study of Table V. All approaches use the same values for 3D localization and orientation, but they differ in their uncertainty

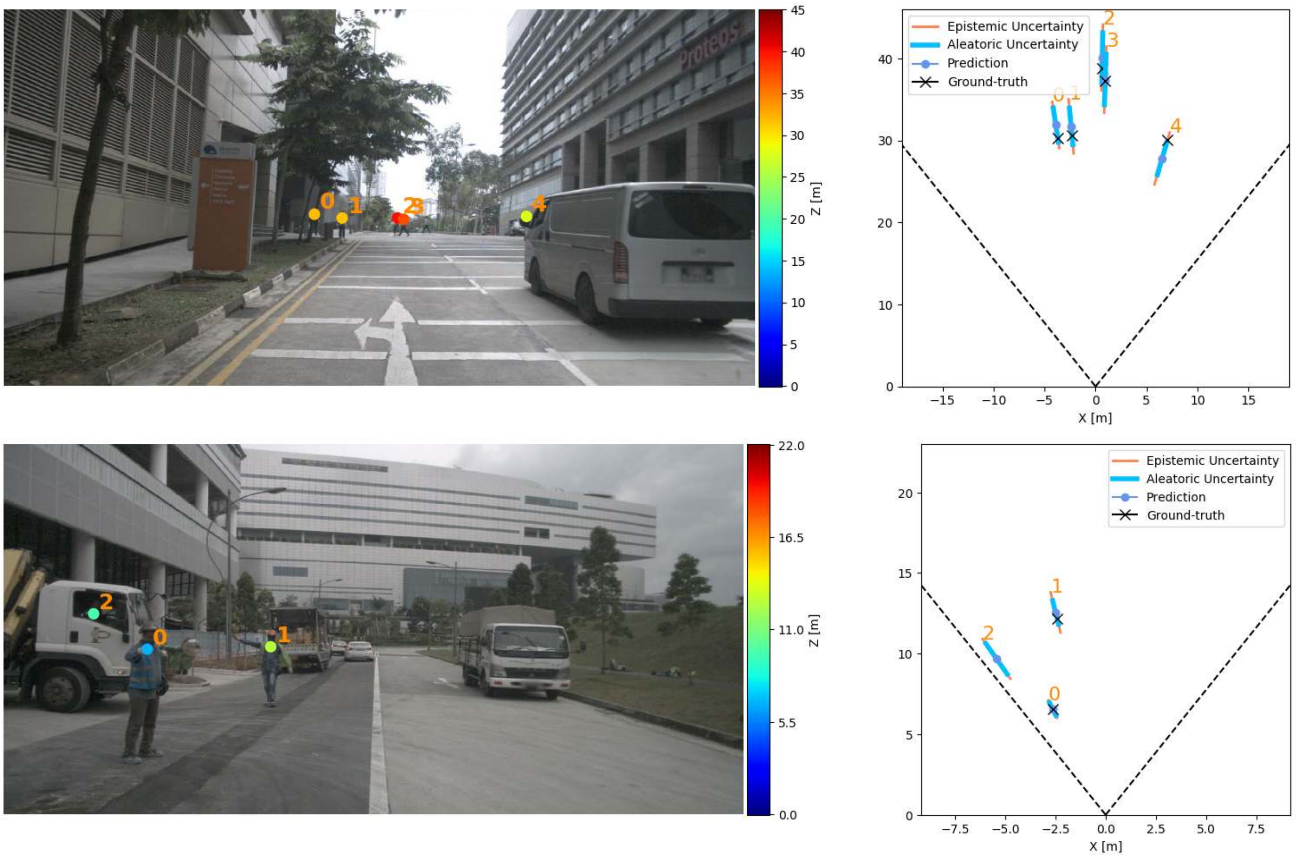

Fig. 9. These examples show 1) why relying on homography or assuming a flat plane can be dangerous, and 2) the importance of uncertainty estimation. In the top image, the road is uphill and the assumption of a constant flat plane would not stand. MonoLoco++ accurately detects people up to 40 meters away. Instance 4 is partially occluded by a van and this is reflected in a higher uncertainty. In the bottom image, we also detect a person inside a truck. No ground-truth is available for the driver but empirically the prediction looks accurate. Furthermore, the estimated uncertainty increases, which is a useful indicator to warn about critical samples.

TABLE V

ACCURACY IN RECOGNIZING THE *talking* ACTIVITY ON THE COLLECTIVE ACTIVITY DATASET [31]. IN ALL CASES THE DISTANCE HAS BEEN ESTIMATED BY OUR MONOLOCO++. "W/O ORIENTATION", DOES NOT USES THE ESTIMATED ORIENTATION, WHILE "DETERMINISTIC" LEVERAGES ORIENTATION BUT NOT THE UNCERTAINTY. "TASK ERROR UNCERTAINTY" REFERS TO THE DISTANCE-BASED UNCERTAINTY DUE TO AMBIGUITY IN THE TASK (EQ. 2), "MONOLOCO++ UNCERTAINTY" REFERS TO THE INSTANCE-BASED UNCERTAINTY ESTIMATED BY OUR MONOLOCO++

Method	Accuracy (%) ↑	Recall (%)↑
W/o Orientation	67.0	97.2
Deterministic	83.7	97.2
Task Error Uncertainty	91.3	97.2
MonoLoco++ Uncertainty	**91.5**	97.2

component. The biggest improvement is given from deterministic approaches (Row 1, Row 2) to a probabilistic one. Row 3 refers to the task error uncertainty of Eq. 2, which grows linearly with distance. Row 4 refers to the estimated confidence interval from MonoLoco++, which are unique for each person. The role of uncertainty is also shown in Figures 8 and 10, where 3D localization errors are compensated by the voting procedure.

D. Social Distancing

Regarding social distancing, there are no fixed rules for evaluation. As previously discussed, the risk of contagion is higher when people are talking to each other [18], yet it may be necessary to maintain social distancing also when people are simply too close. Our goal is not to provide effective rules, but a framework to assess whether a given set of rules is respected.

1) Datasets: In the absence of a dataset for social distancing, we created one by augmenting 3D labels of the KITTI dataset [30]. We apply Eq. 8 using the ground-truth localization and orientation to define whether people are violating social distancing. Once every person is assigned a binary attribute, we evaluate our accuracy on this classification task using our estimated 3D localization and orientation and applying the same set of rules.

2) Evaluation: We evaluate on the augmented KITTI dataset where every person has been assigned a binary attribute for social distancing. Coherently with the monocular 3D localization task, we evaluate on the val split of Chen *et al*. [36] even if no training is performed for this task. We use the same parameters as for the social interaction task, only relaxing the constraint on how people should look inward the o-space, and we set $R_{max} = 2 * r_{o-space}$. This corresponds to verifying whether both candidate centers $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$ are inside the o-space,

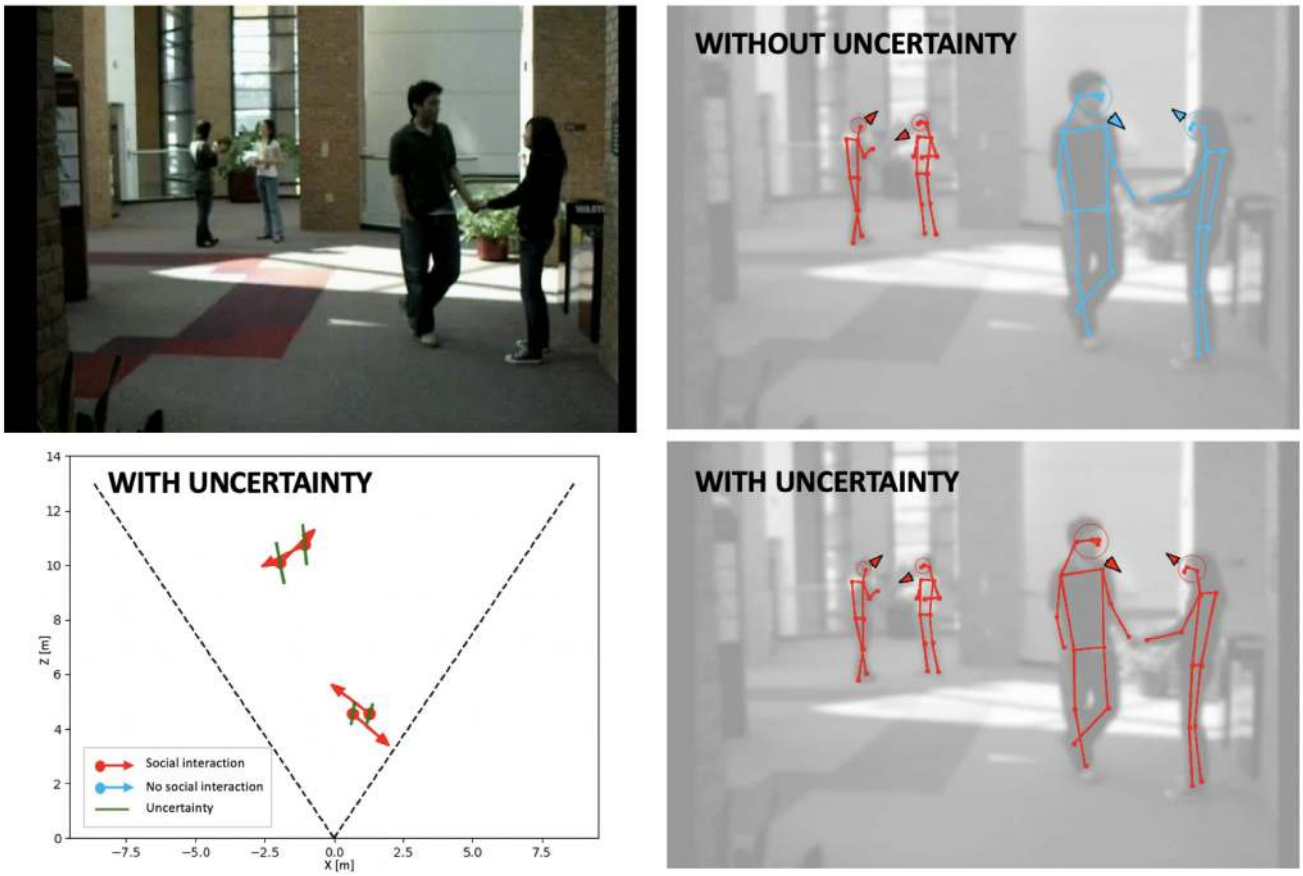

Fig. 10. Estimating whether people are talking to each other(social interaction). Even small errors in 3D localization can lead to wrong predictions. As shown in the bird eye view, the estimated locations of the two people is only slightly off due to the height variation of the subjects. Uncertainty estimation compensates the error due to the ambiguity of the task.

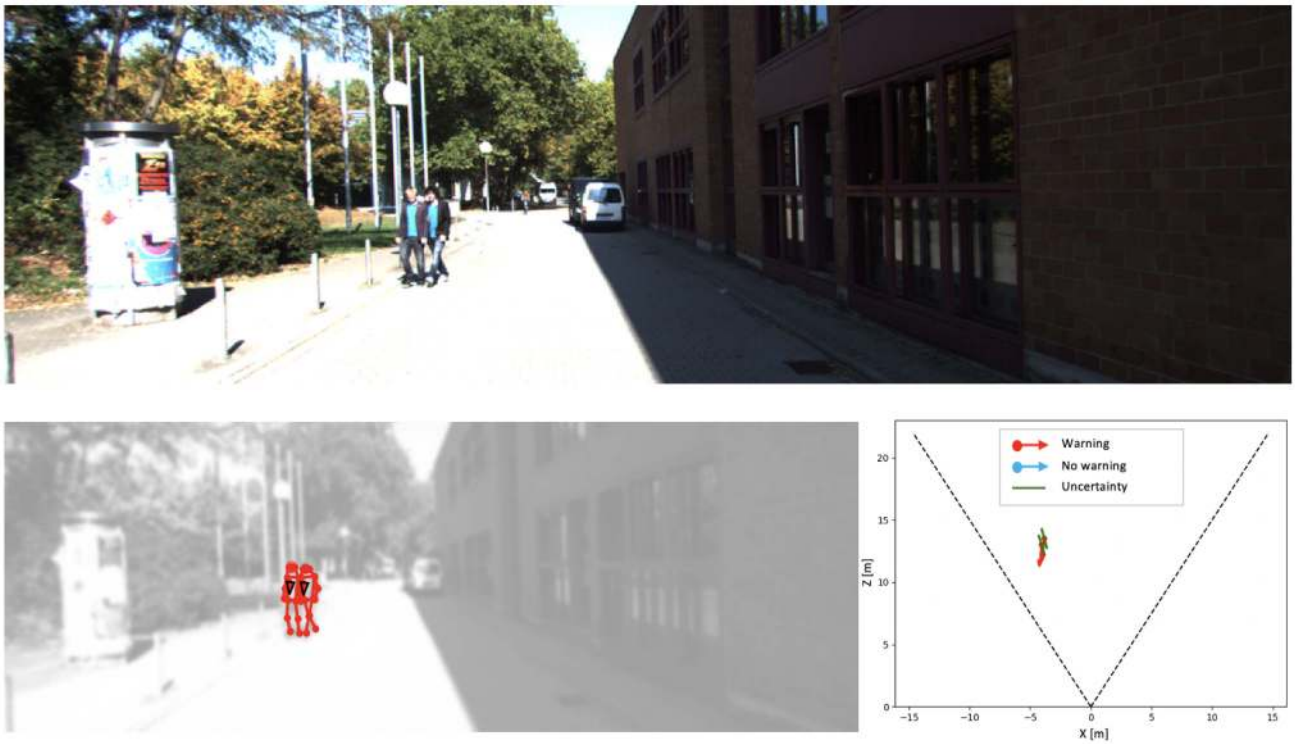

Fig. 11. Qualitative results for the 3D localization task and social distancing. Our MonoLoco++ estimates 3D locations and orientations and raises a warning when social distancing is not respected.

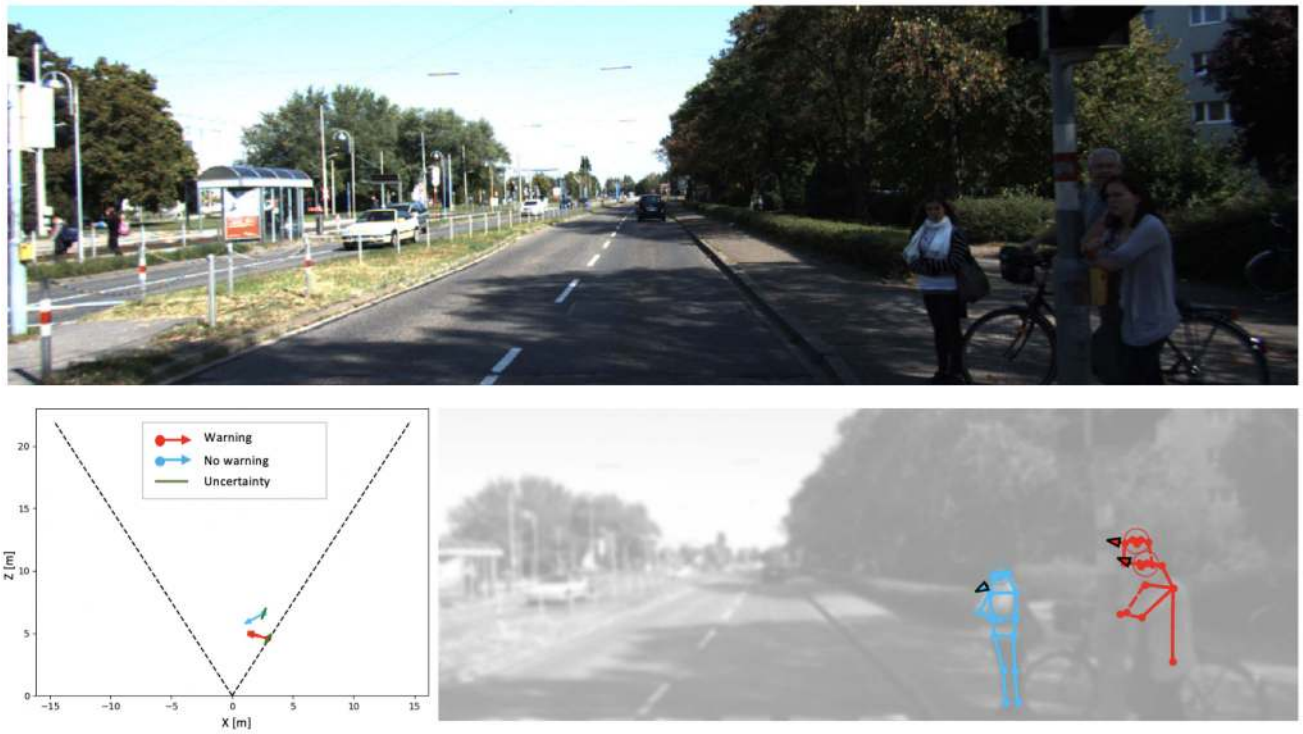

Fig. 12. Qualitative results for the social distancing task in case of three people waiting at the traffic light. Two overlapping people are detected as very close to each other and the system warns for potential risk of contagion. A third person is located slightly more than two meters away and no warning is raised.

as shown in Figure 3. The larger R_{max} in Eq. 8c, the more conservative the social distancing requirement. If Eq. 8c is removed completely, social distancing would only depend on the distance between people.

3) Results: Using the augmented KITTI dataset, we analyze whether social distancing is respected for 1760 people in the validation set. Using the ground-truth localization and orientation we generate labels for which 36.8% of people do not comply with social distancing requirements. This is reasonable as the KITTI dataset contains many crowded scenes. As shown in Table VI, our MonoLoco++ obtains an accuracy of 84.0%. We note that this dataset is more challenging than the Collective Activity one [31], as it includes people 40+ meters away as well as occluded instances. Qualitative results are shown in Figures 11 and 12, where our method estimates 3D localization and orientation, and verify social distancing compliance. In particular, Figure 12 shows that the network is able to accurately localize two overlapping people and recognize a potential risk of contagion, also based on people's relative orientation. In addition, we notice that orientation has a direct impact on reducing false alarms. Without orientation, the network estimates that 43 % of instances violate social distancing requirements. Including orientation, the estimated number reaches 37%, almost on par with the ground-truth value of 38%.

TABLE VI

ACCURACY IN MONITORING SOCIAL DISTANCING ON KITTI DATASET [30]. IN ALL CASES THE DISTANCE HAS BEEN ESTIMATED BY OUR MONOLOCO++. "W/O ORIENTATION", DOES NOT USES ORIENTATION TO ACCOUNT FOR SOCIAL DISTANCING, WHILE "DETERMINISTIC" LEVERAGES ORIENTATION BUT NOT THE UNCERTAINTY. "TASK ERROR U." REFERS TO THE DISTANCE-BASED UNCERTAINTY DUE TO AMBIGUITY IN THE TASK (EQ. 2), "MONOLOCO++ U." REFERS TO THE INSTANCE-BASED UNCERTAINTY ESTIMATED BY OUR MONOLOCO++

Method	Accuracy (%) ↑ [Recall (%) ↑]			
	Easy	*Mod.*	*Hard*	*All*
W/o Orientation	84.0 [95]	80.9 [75]	82.5 [33]	83.3 [75]
Deterministic	80.5 [95]	77.9 [75]	79.0 [33]	79.8 [75]
Task Error U.	84.2 [95]	81.4 [75]	85.3 [33]	83.6 [75]
MonoLoco++ U.	**84.7** [95]	**81.6** [77]	**85.3** [33]	**84.0** [75]

VI. PRIVACY

Our network analyzes 2D poses and does not require any image to process the scene. In fact in Figures 1, 11 and 12, the original image is only shown to clarify the context, but it is not processed directly by MonoLoco++. We leverage an off-the-shelf pose detector which could be embedded in the camera itself. We have designed our system to encourage a privacy-by-design policy [118], where images are processed internally by smart cameras [119] and only 2D poses are sent remotely to a secondary system. The 2D poses do not contain any sensitive data but are informative enough to monitor social distancing.

We also note that smart cameras differ from other technologies by being non-invasive and mostly non-collaborative [118]. Differently from mobile applications, the user is not requested to share any personal data. On the contrary, a low-dimensional representation such as a 2D pose may be challenging for accurate 3D localization, but its ambiguity may prove useful for privacy concerns.

VII. CONCLUSION

We have presented a new deep learning method that perceives humans' 3D locations and their body orientations from

monocular cameras. We emphasized that the main challenge of perceiving social interactions is the ambiguity in 3D localizing people from a single image. Thus, we presented a method that predicts confidence intervals in contrast to point estimates leading to state-of-the-art results. Our system works with a single RGB image, shows cross-dataset generalization properties, and does not require homography calibration, making it suitable for fixed or mobile cameras already installed in transportation systems.

While we have demonstrated the strengths of our method on popular tasks (monocular 3D localization and social interaction recognition), the COVID-19 outbreak has highlighted more than ever the need to perceive humans in 3D in the context of intelligent systems. We argued that to monitor social distancing effectively, we should go beyond a measure of distance. Orientations and relative positions of people strongly influence the risk of contagion, and people talking to each other incur higher risks than simply walking apart. Hence, we have presented an innovative approach to analyze social distancing, not only based on 3D localization but also on social cues. We hope our work will also contribute to the collective effort of preserving people's health while guaranteeing access to transportation hubs.

ACKNOWLEDGMENT

The authors would like to thank their lab members and reviewers for their valuable comments.

REFERENCES

[1] D. F. Llorca *et al.*, "Stereo regions-of-interest selection for pedestrian protection: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 25, pp. 226–237, Dec. 2012.

[2] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *J. Math. Imag. Vis.*, vol. 41, nos. 1–2, pp. 39–58, Sep. 2011.

[3] A. Palffy, J. F. P. Kooij, and D. M. Gavrila, "Occlusion aware sensor fusion for early crossing pedestrian detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1768–1774.

[4] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[5] L. Bertoni, S. Kreiss, and A. Alahi, "MonoLoco: Monocular 3D pedestrian localization and uncertainty estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6861–6871.

[6] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, "TANet: Robust 3d object detection from point clouds with triple attention," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 11677–11684.

[7] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 68–84.

[8] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 641–656.

[9] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253.

[10] A. Alahi, M. Bierlaire, and P. Vandergheynst, "Robust real-time pedestrians detection in urban environments with low-resolution cameras," *Transp. Res. C, Emerg. Technol.*, vol. 39, pp. 113–128, Feb. 2014.

[11] D. Delannay, N. Danhier, and C. De Vleeschouwer, "Detection and recognition of sports (wo) men from multiple views," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug. 2009, pp. 1–7.

[12] Z. Hu, C. Wang, and K. Uchimura, "3D vehicle extraction and tracking from multiple viewpoints for traffic monitoring by using probability fusion map," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2007, pp. 30–35.

[13] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4490–4499.

[14] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 918–927.

[15] W. Bank, "Protecting public transport from the coronavirus and from financial collapse," World Bank Org., Washington, DC, USA, Tech. Rep., 2020.

[16] P. L. Remington, W. N. Hall, I. H. Davis, A. Herald, and R. A. Gunn, "Airborne transmission of measles in a physician's office," *Jama*, vol. 253, no. 11, pp. 1574–1577, 1985.

[17] E. Bromage, "The risks-know them-avoid them," Univ. Massachusetts Dartmouth, Dartmouth, MA, USA, Tech. Rep., 2020.

[18] S. Asadi, A. S. Wexler, C. D. Cappa, S. Barreda, N. M. Bouvier, and W. D. Ristenpart, "Aerosol emission and superemission during human speech increase with voice loudness," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[19] V. Stadnytskyi, C. E. Bax, A. Bax, and P. Anfinrud, "The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 22, pp. 11875–11877, Jun. 2020.

[20] R. J. Rummel, *Understanding Conflict and War: The Just Peace*, vol. 5. Beverly Hills, CA, USA: Sage, 1981.

[21] Y. Zhao, "Mobile phone location determination and its impact on intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 55–64, Mar. 2000.

[22] P. A. Zandbergen, "Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning," *Trans. GIS*, vol. 13, pp. 5–25, Jun. 2009.

[23] P. Kasemsuppakorn and H. A. Karimi, "A pedestrian network construction algorithm based on multiple GPS traces," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 285–300, Jan. 2013.

[24] M. Cristani *et al.*, "Social interaction discovery by statistical analysis of f-formations," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, vol. 2, 2011, p. 4.

[25] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 290–297.

[26] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, "Recognizing proxemics in personal photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3522–3529.

[27] E. S. Aimar, P. Radeva, and M. Dimiccoli, "Social relation recognition in egocentric photostreams," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3227–3231.

[28] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11977–11986.

[29] A. Kendon, *Conducting Interaction: Patterns Behavior Focused Encounters*, vol. 7. Cambridge, U.K.: CUP Archive, 1990.

[30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[31] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Oct. 2009, pp. 1282–1289.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[35] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2017, pp. 1302–1310.

[36] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2147–2156.

[37] W. Deng, L. Bertoni, S. Kreiss, and A. Alahi, "Joint human pose estimation and stereo 3D localization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 2324–2330.

[38] L. Bertoni, S. Kreiss, T. Mordan, and A. Alahi, "MonStereo: When monocular and stereo meet at the tail of 3D human localization," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2021.

[39] A. A. M. Muzahid, W. Wan, F. Sohel, L. Wu, and L. Hou, "CurveNet: Curvature-based multitask learning deep networks for 3D object recognition," *IEEE/CAA J. Automatica Sinica*, early access, Jul. 24, 2020, doi: 10.1109/JAS.2020.1003324.

[40] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[41] A. Alahi *et al.*, "Learning to predict human behavior in crowded scenes," in *Group and Crowd Behavior for Computer Vision*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 183–207.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[43] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[44] C. Sun *et al.*, "Proximity based automatic data annotation for autonomous driving," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 2, pp. 395–404, Mar. 2020.

[45] H. Caesar *et al.*, "NuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*. [Online]. Available: http://arxiv.org/abs/1903.11027

[46] M.-F. Chang *et al.*, "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8748–8757.

[47] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2446–2454.

[48] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," Jul. 2020, *arXiv:2007.03639*. [Online]. Available: http://arxiv.org/abs/2007.03639

[49] E. T. Hall, *The Hidden Dimension*, vol. 609. Garden City, NY, USA: Doubleday, 1966.

[50] G. Papandreou *et al.*, "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3711–3719.

[51] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[52] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[53] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.

[54] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2277–2287.

[55] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 269–286.

[56] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 417–433.

[57] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association," Mar. 2021, *arXiv:2103.02440*. [Online]. Available: http://arxiv.org/abs/2103.02440

[58] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[59] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1561–1570.

[60] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu, "Deep network for the integrated 3D sensing of multiple people in natural images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8410–8419.

[61] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.

[62] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[63] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7122–7131.

[64] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3D object detection leveraging accurate proposals and shape reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11867–11876.

[65] A. Simonelli, S. R. Bulo, L. Porzi, M. L. Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection: From single to multi-class recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 18, 2020, doi: 10.1109/TPAMI.2020.3025077.

[66] Z. Liu, Z. Wu, and R. Toth, "SMOKE: Single-stage monocular 3D object detection via keypoint estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 996–997.

[67] M. E. Kundegorski and T. P. Breckon, "A photogrammetric approach for real-time 3d localization and tracking of pedestrians in monocular infrared imagery," *Proc. SPIE*, vol. 9253, Oct. 2014, Art. no. 92530I.

[68] A. Alahi, A. Haque, and L. Fei-Fei, "RGB-W: When vision meets wireless," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3289–3297.

[69] A. Alahi, M. Bierlaire, and M. Kunt, "Object detection and matching with mobile cameras collaborating with fixed cameras," in *Proc. Workshop Multi-Camera Multi-Modal Sensor Fusion Algorithms Appl. (M2SFA2)*, 2008.

[70] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Tracking millions of humans in crowded spaces," in *Group and Crowd Behavior for Computer Vision*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 115–135.

[71] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7074–7082.

[72] Z. Qin, J. Wang, and Y. Lu, "MonoGRNet: A geometric reasoning network for monocular 3D object localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jan. 2019, pp. 8851–8858.

[73] H.-N. Hu *et al.*, "Joint monocular 3D vehicle detection and tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5390–5399.

[74] B. Xu and Z. Chen, "Multi-level fusion based 3D object detection from monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2345–2353.

[75] F. Manhardt, W. Kehl, and A. Gaidon, "ROI-10D: Monocular lifting of 2D detection to 6D pose and metric shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2069–2078.

[76] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3D object detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019.

[77] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1903–1911.

[78] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 924–933.

[79] F. Chabot, M. 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," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1827–1836.

[80] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-level 3D object reconstruction via render-and-compare," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3559–3568.

[81] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Comput.*, vol. 3, no. 4, pp. 461–483, Dec. 1991.

[82] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. New York, NY, USA: Springer, 2012.

[83] A. Graves, "Practical variational inference for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2348–2356.

[84] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[85] T. Salimans, D. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1218–1226.

[86] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6402–6413.

[87] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[88] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3581–3590.

[89] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[90] J. Postels, F. Ferroni, H. Coskun, N. Navab, and F. Tombari, "Sampling-free epistemic uncertainty estimation using approximated variance propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2931–2940.

[91] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.

[92] J. Mukhoti and Y. Gal, "Evaluating Bayesian deep learning methods for semantic segmentation," 2018, *arXiv:1811.12709*. [Online]. Available: http://arxiv.org/abs/1811.12709

[93] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for Lidar 3D vehicle detection," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3266–3273.

[94] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10873–10883.

[95] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, "Reconfiguring spatial formation arrangement by robot body orientation," in *Proc. 5th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2010, pp. 285–292.

[96] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah, "Social cues in group formation and local interactions for collective activity analysis," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2013, pp. 539–548.

[97] L. Bazzani *et al.*, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Syst.*, vol. 30, no. 2, pp. 115–127, May 2013.

[98] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "A game-theoretic probabilistic approach for detecting conversational groups," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2014, pp. 658–675.

[99] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PLoS ONE*, vol. 10, no. 5, May 2015, Art. no. e0123783.

[100] M. Aghaei, M. Dimiccoli, and P. Radeva, "Towards social interaction detection in egocentric photo-streams," *Proc. SPIE*, vol. 9875, Dec. 2015, Art. no. 987514.

[101] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1868–1877.

[102] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[103] S. Se and M. Brady, "Ground plane estimation, error analysis and applications," *Robot. Auto. Syst.*, vol. 39, no. 2, pp. 59–71, May 2002.

[104] X. Chen *et al.*, "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[105] P. M. Visscher, "Sizing up human height variation," *Nature Genet.*, vol. 40, no. 5, pp. 489–490, May 2008.

[106] J. Freeman, T. Cole, S. Chinn, P. Jones, E. White, and M. Preece, "Cross sectional stature and weight reference curves for the UK, 1990," *Arch. Disease Childhood*, vol. 73, no. 1, pp. 17–24, 1995.

[107] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7644–7652.

[108] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[109] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[110] S. Wirges, M. Reith-Braun, M. Lauer, and C. Stiller, "Capturing object detection uncertainty in multi-layer grid maps," 2019, *arXiv:1901.11284*. [Online]. Available: http://arxiv.org/abs/1901.11284

[111] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[112] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[113] R. Drillis, R. Contini, and M. Bluestein, "Body segment parameters," New York Univ., School Eng. Sci., New York, NY, USA, Tech. Rep., 1969.

[114] H. Han, M. Zhou, and Y. Zhang, "Can virtual samples solve small sample size problem of KISSME in pedestrian re-identification of smart transportation?" *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3766–3776, Sep. 2020.

[115] C. Caetano, F. Bremond, and W. R. Schwartz, "Skeleton image representation for 3D action recognition based on tree structure and reference joints," in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2019, pp. 16–23.

[116] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4315–4324.

[117] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. M. Snoek, "Actor-transformers for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 839–848.

[118] M. Cristani, A. D. Bue, V. Murino, F. Setti, and A. Vinciarelli, "The visual social distancing problem," *IEEE Access*, vol. 8, pp. 126876–126886, 2020.

[119] A. N. Belbachir, *Smart Cameras*, vol. 2. USA: Springer, 2010.

Lorenzo Bertoni received the master's degree from the University of Illinois at Chicago and the master's degree from the Polytechnic University of Turin. He is currently pursuing the Ph.D. degree with the Visual Intelligence for Transportation (VITA) Lab, EPFL, Switzerland, focusing on 3D vision for vulnerable road users. Before joining EPFL, he was a Management Consultant with Oliver Wyman and a Visiting Researcher with the University of California, Berkeley, working on predictive control for autonomous vehicles.

Sven Kreiss is currently a Post-Doctoral Researcher with the Visual Intelligence for Transportation (VITA) Lab, EPFL, Switzerland, focusing on perception with composite fields. Before returning to academia, he was the Senior Data Scientist with Sidewalk Labs (Alphabet, Google sister) and worked on geospatial machine learning for urban environments. Prior to his industry experience, he developed statistical tools and methods used in particle physics research.

Alexandre Alahi received the Ph.D. degree from EPFL. He spent five years at Stanford University as a Post-Doctoral Researcher and a Research Scientist. He is currently an Assistant Professor with EPFL. His research enables machines to perceive the world and make decisions in the context of transportation problems and smart environments. He worked on the theoretical challenges and practical applications of socially-aware Artificial Intelligence, i.e., systems equipped with perception and social intelligence. He awarded the Swiss NSF early and advanced researcher grants for his work on predicting human social behavior. He has also co-founded multiple startups, such as Visiosafe, and won several startup competitions. He was elected as a one of the Top 20 Swiss Venture leaders in 2010.