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 deep learning method that predicts confidence intervals in contrast to point estimates. Our neural network architecture estimates humans 3D body locations and their orientation with a measure of uncertainty. Our vision-based system (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. Indeed, we show that we can 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.

I. INTRODUCTION

Over the past decades, we witnessed new emerging technologies to localize humans in 3D ranging from vision-based solutions [1], [2], [3], [4], [5], to LiDAR-based ones [6], [7] and multi-sensor approaches [8], [9]. On one hand, vision-based technologies can capture detailed body poses and texture properties, but relies on a costly calibrated network of cameras [10], [11], [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 [13], [14], [4]. 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 transports [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 this quest. In fact, talking with a person does not incur the same risk of infection than crossing 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 the exposure time escalates, but the act of speaking itself increases the release of respiratory droplets about 10 fold [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’s that influence or that take account of 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], [22], [23]. 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 outdoor or within large facilities.

In this paper, we propose a deep learning approach that perceive 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 3D location and orientation of people in a scene allows to accurately retrieve 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.
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, which allow to verify social distancing compliance while preserving privacy.

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 at https://github.com/vita-epfl/monstereo.

II. RELATED WORKS

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 revolution 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] or depth estimation [37]. A crucial factor in this transformation has been the release of massive datasets for 2D [38], [39] and 3D tasks [40], [41], [42], 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 [43]. The remaining of this section is organized as follow. 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 included 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.

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 [44], [45], [46], [47]. Bottom-up approaches estimate separately each body joint through convolutional architectures and then combine them to obtain a full human pose [35], [48], [49], [50], [51]. More recently PifPaf [28] proposed a method tailored for autonomous driving scenarios which performs well in low-resolution, crowded and occluded scenes. Related to our work is Simple Baseline [52], which showed the effectiveness of latent information contained in 2D joints stimuli. They achieved state-of-the-art results by simply predicting 3D joints from 2D poses through a light, fully connected network. However, similarly to [53], [54], [55], they estimated relative 3D joint positions, not providing any information about the real 3D location in the scene.

Monocular 3D Object Detection. Recent approaches for monocular 3D object detection in the transportation domain
focused only on vehicles as they are rigid objects with known shape. To the best of our knowledge, only the very recent Mono3D [56] evaluated pedestrians from monocular RGB images, leveraging point clouds at training time to learn local shapes of objects. Kundegorski and Breckon [57] achieved reasonable performances combining infrared imagery and real-time photogrammetry. Alahi et al. combined monocular images with wireless signals [58] or with additional visual priors [59], [10]. The seminal work of Mono3D [56] exploited deep learning to create 3D object proposals for car, pedestrian and cyclist categories but it did not evaluate 3D localization of pedestrians. It assumed a fixed ground plane orthogonal to the camera and the proposals were then scored based on scene priors, such as shape, semantic and instance segmentations. Following methods continued to leverage Convolutional Neural Networks and focused only on Car instances. To regress 3D pose parameters from 2D detections, Deep3DBox [60]. MonoGRnet [61], and Hu et al. [62] used geometrical reasoning for 3D localization, while Multi-fusion [63] and ROI-10D [64] incorporated a module for depth estimation. Recently, Roddick et al. [65] escaped the image domain by mapping image-based features into a birds-eye view representation using integral images. Another line of work fits 3D templates of cars to the image [66], [67], [68], [69]. While many of the related methods achieved reasonable performances for vehicles, current literature lacks monocular methods addressing other categories in the context of autonomous driving, such as pedestrians and cyclists.

Uncertainty Estimation in Computer Vision. Deep neural networks need to have the ability not only to provide the correct outputs but also a measure of uncertainty, especially in safety-critical scenarios like autonomous driving. Traditionally, Bayesian Neural Networks [70], [71] were used to model epistemic uncertainty through probability distributions over the model parameters. However, these distributions are often intractable and researchers have proposed interesting solutions to perform approximate Bayesian inference to measure uncertainty, including Variational Inference [72], [73], [74] and Deep Ensembles [75]. Alternatively, Gal et al. [76], [77] showed that applying dropout [78] at inference time yields a form of variational inference where parameters of the network are modeled as a mixture of multivariate Gaussian distributions with small variances. This technique, called Monte Carlo (MC) dropout, became popular also due to its adaptability to non-probabilistic deep learning frameworks. Very recently, Postels et al. [79] 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 [80], [79], scene segmentation [81], [80] and, more recently, LiDAR 3D object detection for cars [82].

B. Social Interactions

In this work, we aim to capture social interactions among people and monitor social distancing from visual cues. Related works include the broad field of behavioral science [83], but here we focus on the subfield called *proxemics*, which investigates how people use and organize the space they share with others [43], [25]. 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 to each other. They are characterized by three types of social spaces [43], [25]:

1) o-space: an empty space that surround the people and preserve the intimacy of people. 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-formation are visual, L-shape, and side-by-side. For larger groups, a circular formation is typically formed [84]. An example of 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] studied how people get closer to each other when the social relation is more intimate. Following works have proposed various techniques to automatically detect F-formations in heterogeneous real crowded scenarios [85], [86], [87], [88]. In all approaches is evident how the detection of F-formations is critical to inference social relations and we decided to follow their lead. This line of works, however, considers as input the position of people on the ground floor and their orientation [25] or requires an homography estimation to compute the x-y-z coordinates of pedestrians [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.

Social interactions have been also studied in the context of personal photos [26] or egocentric photo-stream [89], [27] Both approaches assumes humans to stand less than few meters apart from each other and the camera, and do not scale to large range applications, such as monitoring an airport terminal. Recently, deep learning approach has been adopted to understand social interactions under a different perspective. Joo et al. [83] learned to predict behavioral cues of a target person (e.g., body orientation) from the position and orientation of another person. They aimed to 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 pedestrians from monocular images is a fundamentally ill-posed problem. The majority of previous works has circumvented this challenge by assuming a planar ground plane and estimating an homography by manual measurement or by knowing some reference elements [90], [24], [56], [91]. This approach only works if everyone is on the same plane (what if people are descending stairs?) and requires a new calibration for each environment. Therefore in this work, we aim for a more direct approach: directly estimating distance of pedestrians without relying on contextual cues, such as scene geometry. 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 quest 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 [57], we model the localization error due to variation of height 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, $d_{h-mean}/h_{mean} = d_{gt}/h_{gt}$, 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 $d_{h-mean}$ the estimated distance under the $h_{mean}$ assumption. We can define the task error for any person instance in the dataset as:

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

(1)

Previous studies from a population of 63,000 European adults have shown that the average height is 178cm for males and 165cm for females with a standard deviation of around 7cm in both cases [92]. 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 [93], 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

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

(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% [93], [57]. 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 [57]. Figure 4a 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 pedestrians, while imposing an intrinsic limit, does not prevent from robust localization in general cases.

IV. Method

The goals of our method are (i) to detect pedestrians 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 joints are a meaningful low-level representation which provides invariance to many factors, including background scenes, lighting, textures and clothes. Second, we use the 2D joints as input to a feed-forward neural network which predicts 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 [43] and recognize social interactions.

A. 3D Human detections

The task of 3D object detection is defined as detecting 3D location of objects along with their orientation and dimensions [30], [40]. 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 without supervision and predicts confidence intervals in contrast to point estimates. The task error modeled in Eq. 2 allows to compare the predicted confidence intervals with the intrinsic ambiguity of the task.

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

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

(3)

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

2D Human Poses. We obtain 2D joint locations of pedestrians using the off-the-shelf pose detectors PifPaf [28], 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.

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...
Fig. 2: Network 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 pedestrian and the localization uncertainty. 3D location is estimated with spherical coordinates: $r$, azimuthal angle $\beta$, and polar angle $\psi$. Every fully connected layer (FC) outputs 1024 features and is followed by a Batch Normalization layer (BN) \cite{94} and a ReLU activation function. **Social interactions/distancing**: estimates from MonoLoco++ are analyzed with an all-vs-all approach to discover F-formations using Eq. 8. 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 $(r, \beta, \psi)$, namely radial distance $r$, azimuthal angle $\beta$, and polar angle $\psi$. 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 $r$ and not on its depth $z$ \cite{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 \cite{95}, the viewpoint angle is not equal to the object orientation as people at different locations may share the same orientation $\theta$ but results in different projections. Hence, we predict the viewpoint angle $\alpha$, which is defined as $\alpha = \theta + \beta$, where $\beta$ denotes the azimuth of the pedestrian with respect to the camera. Similarly to \cite{95}, we also parametrize 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. At last, we profit from aleatoric uncertainty \cite{80}, i.e., the uncertainty of each task, to weigh our loss function for our multi-task learning. Our minimization objective for our multi-output model follows the formulation in \cite{96}.

**Base Network.** The building blocks of our model are shown in Figure 2. The architecture, inspired by Martinez et al. \cite{52}, is a simple, deep, fully-connected network with six linear layers with 256 output features. It includes dropout \cite{78} after every fully connected layer, batch-normalization \cite{94} and residual connections \cite{97}. The model contains approximately 400k training parameters.

**MonoLoco++ vs MonoLoco.** We refer to our method as MonoLoco++. Technically, it differs from the previous MonoLoco \cite{5} by:

- the multi-task approach to combine 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 road for activity recognition and social distancing, which was not possible using MonoLoco \cite{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 technical improvements also 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* \cite{98}, \cite{80}. Aleatoric uncertainty is an intrinsic property of the task and the inputs. It does not decrease when collecting more data. In the context of 3D monocular localization, the intrinsic ambiguity of the task represents a quota of aleatoric uncertainty. In addition, some inputs may be more noisy than others, leading to an input-dependent aleatoric uncertainty. Epistemic uncertainty is a property of the model parameters, and it can be reduced by gathering more data. It is useful to quantify the ignorance of the model about the collected data, e.g., in case of out-of-distribution samples.

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

\[
L_{\text{Laplace}}(x|d, b) = \frac{1 - d/x}{b} + \log(2b),
\]

where \( x \) represents the ground-truth distance, and \( d, b \) the predicted distance and the spread, making this training objective an attenuated \( L_1 \)-type loss via spread \( b \). During training, the model has the freedom to predict a large spread \( b \), leading to attenuated gradients for noisy data. The uncertainty is estimated in an unsupervised way, since no supervision is provided. At inference time, the model predicts the distance \( d \) and a spread \( b \) which indicates its confidence about the predicted distance. Following [80], 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 [80], [99], we design a Laplace loss which works with relative distances to keep 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.

**Epistemic Uncertainty.** To model epistemic uncertainty, we follow [76], [80] and consider each parameter as a mixture of two multivariate Gaussians with small variances and means 0 and \( \theta \). The additional minimization objective for \( N \) data points is:

\[
L_{\text{dropout}}(\theta, p_{\text{dropout}}) = \frac{1 - p_{\text{dropout}}}{2N} ||\theta||^2.
\]

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 \( p_{\text{dropout}} \) 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:

\[
\text{Var}(\tilde{X}) = \frac{1}{T^2} \sum_{t=1}^T \sum_{i=1}^I \tilde{x}^2_{t, i}(d_t, b_t) = \frac{1}{T^2} \sum_{t=1}^T \sum_{i=1}^I \tilde{x}^2_{t, i}(d_t, b_t)
\]

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 from considering its radius \( r_o \) dynamically as the minimum distance between the center and one of the two people. An illustration of the differences is show in Figure 4. Therefore, given the location of two people in the xz plane \( x \) and their body orientation \( \theta \), we define the center and the radius of the o-space as:

\[
O_{01} = \frac{\mu_0 + \mu_1}{2},
\]

\[
r_{o-space} = \min(|O_{01} - x_o|, |O_{01} - x_1|),
\]

where \( O_{01}, r_{o-space} \) are the center and radius of the resulting o-space. \( \mu_0, \mu_1 \) indicate the location of the two candidate centers of the o-space. In general, \( \mu = [x + r \cos(\theta), z + r \sin(\theta)] \) and is parametrized by the candidate radius \( r \), which depends on the type of relation (intimate, personal, business, etc.) [43].

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

\[
\begin{align*}
(a) & \quad |x_o - x_1| < D_{\text{max}} \\
(b) & \quad |O_{01} - x_i| < r_{o-space} \quad \forall i \neq 0, 1 \\
(c) & \quad |\mu_0 - \mu_1| < R_{\text{max}}
\end{align*}
\]

where \( D_{\text{max}}, R_{\text{max}} \) are the maximum distance 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) defines whether two people stand closer than a maximum distance \( D_{\text{max}} \), hence they lie inside an r-space
(b) verifies the presence of an empty o-space not occupied by any other person (no-intrusion condition)
(c) verifies whether the two people are looking inwards the o-space

We note that condition (c) is empirical as looking inwards is a generic requirement. Two people normally 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 how effective simple rules can be when combined with estimated 3D localization and orientation. We consider two people as interacting to 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.

**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 [19], [17]. Therefore, recognizing social interactions lets the system only warn the people that incur in 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 thorough distance estimation but also leveraging social cues.

**Uncertainty for social interactions.** A deterministic approach can be very sensitive to small error in 3D localization, which we know are inevitable due to the perspective projection. Therefore, we introduce a probabilistic approach that leverage 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.
for KITTI and 14500 for nuScenes teaser. KITTI images are upsampled 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 run the training procedure for 200 epochs using Adam optimizer [100], a learning rate of $10^{-3}$ and mini-batches of 512. The code, available online, is developed using PyTorch [101]. Working with a low-dimensional latent 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 GPU GTX1080Ti.

**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 used for the evaluation. In contrast, analyzing ALE and recall allows for simple but fair comparison. Following KITTI guidelines, we assign to each detected 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.

**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 $D = [x_c, y_c, z_c]^T$ we can back-project each keypoint from the image plane to its 3D location and measure the height of each segment using Eq. [5]. We calculate the mean and the standard deviation in meters of each of the segments for all the instances in the training set. The standard deviation is used to choose the most stable segment for our calculations. For instance, the position of the head with respect to shoulders may vary a lot for each instance. To take into account noise in the 2D joints predictions we also average between left and right keypoints values. The result is a single height $\Delta y_{1-2}$ which represents the average length of two body parts. In practice, our geometric baseline uses the shoulder-hip segment and predicts an average height of 50.5cm. Combining the study on human heights [92] described in Section 3 with the anthropometry study of Drillis et al. [102], we can compare our estimated $\Delta y_{1-2}$ with the human average shoulder-hip height: $0.288 \times 171.5cm = 49.3cm$.

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 $\Delta y_{1-2}$ 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.

**Other baselines.** We compare our monocular method on KITTI against three 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].
- **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** [50] a very recent monocular 3D object detector which leverages point cloud at training time to learn shapes of objects. In contrary, our method does not use any privileged signal at training time.
- **3DOP** [91] is a stereo approach for pedestrians, cars and cyclists and their 3D detections are publicly available.

Finally, in Figure 4a, 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**

**Localization accuracy.** Table [4] summarizes our quantitative results on KITTI. We strongly outperform all the other monocular approaches on all metrics with any of the two models trained either on KITTI or nuScenes. We obtain comparable results with the stereo approach 3DOP [91], which has been trained and evaluated on KITTI and makes use of stereo images during training and test time.

In Figure 4a, 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, with a quasi-linear behaviour which almost replicates the target threshold. Figure 5 and 6 shows qualitative results on challenging images from KITTI and nuScenes datasets, respectively.
TABLE I: Comparing our proposed method against baseline results on 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. K stands for trained on KITTI [30], N for trained on nuScenes teaser [40]. In both cases the evaluation protocol is the same. The model trained on nuScenes shows cross-dataset generalization by obtaining the best results among the monocular methods in the ALE metric.

| Method           | ALE (m) \(\downarrow\) | [Recall (%) \(\uparrow\)] | ALA (%) \(\uparrow\) |
|------------------|-------------------------|-----------------------------|------------------------|
| Mono3D [36]      | 2.28 [98%]              | 3.00 [85%]                  | 51.14 [72%]            |
| MonoPSR [55]     | 0.89 [98%]              | 1.97 [85%]                  | 2.15 [21%]             |
| MonoLoco++ (K) [5] | 0.76 [98%]              | 1.26 [49%]                  | 88.3 [63%]             |
| MonoLoco++ (N) [5] | 0.59 [92%]              | 1.17 [75%]                  | 1.44 [32%]             |
| Our MonoLoco++ (K) | 0.77 [91%]              | 0.88 [75%]                  | 0.85 [71%]             |
| Our MonoLoco++ (N) | 0.72 [91%]              | 0.84 [72%]                  | 0.89 [23%]             |

TABLE II: Precision and recall of uncertainty for KITTI validation set with 50 stochastic forward passes. \(|x - d|\) is the localization error, \(\sigma\) 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.

| Method          | \(|x - d|/\sigma\) | \(|\sigma - e|\) [m] | Recall [%] |
|-----------------|------------------|-----------------------|------------|
| \(p_{\text{drop}} = 0.05\) | 0.60             | 0.90                  | 82.8       |
| \(p_{\text{drop}} = 0.2\)  | 0.58             | 0.96                  | 84.3       |
| \(p_{\text{drop}} = 0.4\)  | 0.50             | 1.26                  | 88.3       |

Aleatoric uncertainty. We compare in Figure 4b the aleatoric uncertainty predicted by our network through spread \(b\) with the task error due to human height variation defined in Eq. 2. The predicted spread \(b\) is a property of each set of inputs and, differently from \(\sigma\), is not only a function of the distance from the camera \(d\). Indeed, the predicted aleatoric uncertainty includes not only the uncertainty due to the ambiguity of the task but also the uncertainty due to noisy observations [80]. i.e., the 2D joints inferred by the pose detector. Hence, we can approximately define the predictive aleatoric uncertainty due to noisy joints as \(b - \hat{e}\) and we observe that the further a person is from the camera, the higher is the term \(b - \hat{e}\). The spread \(b\) is the result of a probabilistic interpretation of the model and the resulting confidence intervals are calibrated. On KITTI validation set they include 68% of the instances.

TABLE III: Impact of different loss functions with Mask R-CNN [40] and PifPaf [28] pose detectors on nuScenes teaser validation set [40].

| Mask R-CNN \([40]\) | ALE [m] | \(|x - d|\) / \(\sigma\) | \(|\sigma - e|\) [m] |
|---------------------|---------|------------------------|-----------------------|
| Geometric           | 0.79    | 1.52                   | 3.17                  |
| Gaussian loss       | 0.90    | 1.28                   | 2.34                  |
| Laplace Loss        | 0.74    | 1.17                   | 2.25                  |
| PifPaf [28]         |         |                        |                       |
| Geometric           | 0.83    | 1.40                   | 2.15                  |
| Gaussian Loss       | 0.89    | 1.22                   | 2.14                  |
| Laplace Loss        | 0.75    | 1.19                   | 2.24                  |

Combined uncertainty. The combined aleatoric and epistemic uncertainties are captured by sampling from multiple Laplace distributions using MC dropout. The magnitude of the uncertainty depends on the chosen dropout probability \(p_{\text{drop}}\) in Eq. 5. In Table 11 we analyze the precision/recall trade-off for different dropout probabilities and choose \(p_{\text{drop}} = 0.2\). We perform 50 computationally expensive forward passes and, for each of them, 100 computationally cheap samples from Laplace distribution using Eq. 6. As a result, 84% of pedestrians lie inside the predicted confidence intervals for the validation set of KITTI.

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.

TABLE IV: Single-image inference time on a GTX 1080Ti for KITTI dataset [30] with PifPaf [28] as pose detector. We only considered images with positive detections. Most computation comes from the pose detector (ResNet 50 / ResNet 152 backbones). For Mono3D, 3DOP and MonoPSR we report published statistics on a Titan X GPU. In the last line, we calculated epistemic uncertainty through 50 sequential forward passes. In future work, this computation can be paralleled.

| Method            | Time [ms] | pose | preprocess | total |
|-------------------|-----------|------|------------|-------|
| Mono3D [36]       | -         | 1800 | 1800       |
| 3DOP [9]          | -         | 2000 | 2000       |
| MonoPSR [55]      | -         | 2000 | 2000       |
| Our MonoLoco++ (1 forward pass) | 89 / 162 | 10 | 99 / 172 |
| Our MonoLoco++ (50 forward passes) | 89 / 162 | 51 | 140 / 213 |

Challenging cases. We qualitatively analyze the role of the predicted uncertainty in case of an outliers in Figure 8. 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], [91]. The road in Figure 8 (top) is uphill 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.

**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. $L_1$-type losses perform slightly better than the Gaussian loss, but the main improvement is given by choosing PifPaf as pose detector.

**Run time.** A run time comparison is shown in Table IV. Our method is faster or comparable to all the other methods, achieving real-time performances.

**C. Social Interactions**

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

**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 to test our 3D localization performances in different scenarios. Compared to other deep learning methods [103], [104], [105], 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.

**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: $D_{\text{max}} = 2$ as maximum distance, and $r_1 = 0.3$, $r_2 = 0.5$, $r_3 = 1$ as radii for o-space candidates. These choices reflect the average distances of intimate rela-

| Method                          | Accuracy (%) ↑ | Recall (%)↑ |
|--------------------------------|---------------|------------|
| Deterministic                  | 80.7          | 96.7       |
| Task Error Uncertainty         | 89.8          | 96.7       |
| MonoLoco++ Uncertainty         | 91.4          | 96.7       |

**TABLE V: Accuracy in recognizing the talking activity on the Collective Activity dataset [31].** The deterministic approach does not leverage 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 (%) ↑] |
|-------------|---------------|---------------|
| Deterministic | 80.5 [95]     | 81.1 [82]     |
| Task Error U. | 83.1 [95]     | 82.9 [82]     |
| MonoLoco++ U. | 83.2 [95]     | 84.1 [82]     |

**TABLE VI: Accuracy in monitoring social distancing on KITTI dataset [30].** The deterministic approach does not leverage uncertainty (U.). **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++.

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 $R_{\text{max}} = r_{o-space}$ 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 analyzes sequences that contain at least a person talking in one of its frames. Consequently, we evaluate a total of 4328 instances, of which 52.8% is talking.

**Voting procedure.** To account for noise in 3D localization, we sample our results from the estimated Laplace distribution parametrized by distance $d$ and spread $b$ (Eq. 4). Each sample vote for a candidate center $\mu$ 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 8. We compare this technique to (i) a deterministic approach by only using the distance $d$, and (ii) a probabilistic approach where the uncertainty is provided by the task error defined in Eq. 2.

**Results.** Table VII 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 91.4% 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 VII. All approaches use the same values for 3D localization and orientation, but they differ in their uncertainty component. The biggest improvement is given from a deterministic approach to a probabilistic one. Row 2 refers to the task error uncertainty of Eq. 2 which grows linearly with distance. Rows 3 refers to the estimated confidence interval from MonoLoco++, which are unique for each person. The role of uncertainty is also shown in Figures 7 and 9 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.

Datasets. In the absence of a dataset for social distancing, we created one by augmenting 3D labels of 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.

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 \( R_{max} = 2 \times r_{o-space} \). This corresponds to verifying whether both candidate centers \( \mu_0, \mu_1 \) are inside the o-space, 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.

Results. Using the augmented KITTI dataset, we analyze whether social distancing is respected for 1760 people. 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 KITTI dataset contains many crowded scenes. As shown in Table VI our MonoLoco++ obtains an accuracy of 83.2%. We note that this dataset is more challenging than the Collective Activity one [31], as it includes people 40+ meters far as well as occluded instances. Qualitative results are shown in Figures 10 and 11 where our method estimates 3D localization and orientation, and verify social distancing compliance. In particular, Figure 11 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.

VI. Privacy

Our network analyzes 2D poses and does not require any image to process the scene. In fact in Figures 1, 10 and 11 the original image is only shown to clarify the context, but is not process 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 [106], where images are processed internally by smart cameras [107] and only 2D poses are sent remotely to a secondary system. The 2D poses do not contain any sensible data but are informative enough to monitor social distancing. We also note that smart cameras differentiate from other technologies by being non-invasive and mostly non-collaborative [106]. 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. Conclusions

We have presented a new deep learning method that perceive humans’ 3D location and their body orientation 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 and does not require homography calibration, making it suited 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 effectively monitoring social distancing, we should go beyond a measure of distance. Orientation and relative positions of people strongly influence the risk of contagion, and people talking to each other incur in 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.
Fig. 5: Illustration of results from 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 matches a ground-truth are shown for clarity.
Fig. 6: 3D localization task. Illustration of results from nuScenes dataset [40] containing true and inferred distance information as well as confidence intervals.
Fig. 7: Estimating whether people are talking to each other. 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 location and orientation of all the people. The color of the arrows indicates whether people are involved in talking.
Fig. 8: 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 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 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, a useful indicator to warn about critical samples.
Fig. 9: Estimating whether people are talking. Even small errors in 3D localization can lead to wrong predictions. As shown in the bird eye view, the estimated location 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. 10: 3D localization task. Illustration of two people walking and talking together. Our MonoLoco++ estimates 3D location, orientation and raises a warning when social distancing is not respected.
Fig. 11: 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.
REFERENCES

[1] D. Llorca, M. Sotelo, A. Hellín, A. Orellana, M. Gavilán, I. Daza, and A. Lorente, “Stereo regions-of-interest selection for pedestrian protection: A survey,” Transportation research part C: emerging technologies, vol. 25, pp. 226–237, 2012.

[2] A. Alahi, L. Jacques, Y. Boursier, and P. Vanderheynst, “Sparsity driven people localization with a heterogeneous network of cameras,” Journal of Mathematical Imaging and Vision, vol. 41, no. 1-2, pp. 39–58, 2011.

[3] A. Palffy, J. F. Kooij, and D. M. Gavrila, “Occlusion aware sensor fusion for early crossing pedestrian detection,” in 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019, pp. 1768–1774.

[4] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzaktis, “A survey on 3d object detection methods for autonomous driving applications,” IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 10, pp. 3782–3795, 2019.

[5] L. Bertoni, S. Kreiss, and A. Alahi, “Monoloco: Monocular 3d pedestrian localization and uncertainty estimation,” in The IEEE International Conference on Computer Vision (ICCV), October 2019.

[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 AAAI Conference on Artificial Intelligence, 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 The European Conference on Computer Vision (ECCV), 2020.

[8] M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep continuous fusion for multi-sensor 3d object detection,” in The European Conference on Computer Vision (ECCV), 2018, pp. 641–656.

[9] D. Xu, D. Anguelov, and A. Jain, “Pointfusion: Deep sensor fusion for 3d bounding box estimation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 244–253.

[10] A. Alahi, M. Bierlaire, and P. Vanderheynst, “Robust real-time pedestrians detection in urban environments with low-resolution cameras,” Transportation research part C: emerging technologies, vol. 39, pp. 113–128, 2014.

[11] D. Delannay, N. Danhier, and C. De Vleeschouwer, “Detection and recognition of sports (wo) men from multiple views,” in 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC). IEEE, 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 2007 IEEE Intelligent Transportation Systems Conference. IEEE, 2007, pp. 30–35.

[13] Y. Zhou and O. Tuzel, “Voxelnet: End-to-end learning for point cloud based 3d object detection,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 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 The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 918–927.

[15] W. Bank, “Protecting public transport from the coronavirus... and from financial collapse,” 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,” 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,” Scientific reports, vol. 9, no. 1, pp. 1–10, 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,” Proceedings of the National Academy of Sciences, 2020.

[20] R. J. Rummel, “Understanding conflict and war: Vol. 5: The just peace,” Beverly Hills, California: Sage Publications, 1981.

[21] Y. Zhao, “Mobile phone location determination and its impact on intelligent transportation systems,” IEEE Transactions on intelligent transportation systems, vol. 1, no. 1, pp. 55–64, 2000.

[22] P. A. Zandbergen, “Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning,” Transactions in GIS, vol. 13, pp. 5–25, 2009.

[23] P. Kasempuppakorn and A. H. Karimi, “A pedestrian network construction algorithm based on multiple gps traces,” Transportation research part C: emerging technologies, vol. 26, pp. 285–300, 2013.

[24] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, “Social interaction discovery by statistical analysis of f-formations,” in British Machine Vision Conference (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 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE, 2011, pp. 290–297.

[26] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, “Recognizing proxemics in personal photos,” in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 3522–3529.

[27] E. S. Aimer, P. Radeva, and M. Dimiccoli, “Social relation recognition in egocentric photostreams,” in The IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 3227–3231.

[28] S. Kreiss, L. Bertoni, and A. Alahi, “Pifpaf: Composite fields for human pose estimation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11977–11986.

[29] A. Kendon, Conducting interaction: Patterns of behavior in focused encounters. CUP Archive, 1990, vol. 7.

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

[31] W. Choi, K. Shahid, and S. Savarese, “What are they doing?: Collective activity classification using spatio-temporal relationship among people,” in 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. IEEE, 2009, pp. 1282–1289.

[32] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444, 2015.

[33] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[80] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” in Advances in Neural Information Processing Systems, 2017, pp. 5574–5584.

[81] J. Mukhoti and Y. Gal, “Evaluating bayesian deep learning methods for semantic segmentation,” arXiv preprint arXiv:1811.12709, 2018.

[82] D. Feng, L. Rosenbaum, and K. Dietmayer, “Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection,” in the IEEE International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 3266–3273.

[83] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, “Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction,” in The IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 753–762.

[84] H. Kuzumoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, “Reconfiguring spatial formation arrangement by robot body orientation,” in 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2010, pp. 285–292.

[85] 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 International Conference on Computer Vision Theory and Applications (VISAPP, 2013, pp. 539–548.

[86] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, “Social interactions by visual focus of attention in a three-dimensional environment,” Expert Systems, Vol. 30, no. 2, pp. 115–127, 2013.

[87] S. Vascon, E. Z. Mequinit, M. Cristani, H. Hung, M. Pelillo, and V. Murino, “A game-theoretic probabilistic approach for detecting conversational groups,” in Asian Conference in Computer Vision (ACCV). Springer, 2014, pp. 658–675.

[88] F. Setti, C. Russell, C. Bassetti, and M. Cristani, “F-detection: Individuating free-standing conversational groups in images,” PloS one, vol. 10, no. 5, p. e0127383, 2015.

[89] M. Aghai, M. Dimiccoli, and P. Radeva, “Towards social interaction detection in egocentric photo-streams,” in Eighth International Conference on Machine Vision (ICMV 2015), vol. 9875. International Society for Optics and Photonics, 2015, p. 987514.

[90] S. Se and M. Brady, “Ground plane estimation, error analysis and applications,” Robotics and Autonomous systems, vol. 39, no. 2, pp. 59–71, 2002.

[91] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fieler, and R. Urtasun, “3d object proposals for accurate object class detection,” in Advances in Neural Information Processing Systems, 2015, pp. 424–432.

[92] P. M. Visscher, “Sizing up human height variation,” Nature genetics, vol. 40, no. 5, p. 489, 2008.

[93] C. Freeman, T. Cole, S. Chinn, P. Jones, E. White, and M. Preece, “Cross sectional stature and weight reference curves for the uk," 1990. Archives of disease in childhood, vol. 73, no. 1, pp. 17–24, 1995.

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

[95] P. Li, X. Chen, and S. Shen, “Stereo r-cnn based 3d object detection for autonomous driving,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7644–7652.

[96] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 7482–7491.

[97] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[98] A. Der Kiureghian and O. Ditlevsen, “Aleatory or epistemic? does it matter?” Structural Safety, vol. 31, no. 2, pp. 105–112, 2009.

[99] W. Girges, M. Reith-Braun, M. Lauer, and C. Stiller, “Capturing object detection uncertainty in multi-layer grid maps,” arXiv preprint arXiv:1901.11284, 2019.

[100] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[101] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., “Pytorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[102] P. Nardini, R. Contini, and M. Buehler, “Body segment parameters,” New York University, School of Engineering and Science, 1969.

[103] C. Caetano, F. Brémont, and W. R. Schwartz, “Skeleton image representation for 3d action recognition based on tree structure and reference joints,” in 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2019, pp. 16–23.

[104] T. Bagautdinov, A. Alahi, F. Fleuret, F. Fua, and S. Savarese, “Social scene understanding: End-to-end multi-person action localization and collective activity recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4315–4324.

[105] K. Evstatiev, R. Sanford, M. Javan, and C. G. Snoek, “Actor-transformers for group activity recognition,” in the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 839–848.

[106] 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.

[107] A. N. Belbachir, Smart cameras. Springer, 2010, vol. 2.

Lorenzo Bertoni is a doctoral student at the Visual Intelligence for Transportation (VITA) lab at EPFL in Switzerland focusing on 3D vision for vulnerable road users. Before joining EPFL, Lorenzo was a visiting researcher at the University of California, Berkeley, working on predictive control of autonomous vehicles. Lorenzo received Bachelors and Masters Degrees in Engineering from the Polytechnic University of Turin and the University of Illinois at Chicago.

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

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