IntrApose

Monocular Driver 6 DOF Head Pose Estimation Leveraging Camera Intrinsics

Roth, Markus; Gavrila, Dariu M.

DOI
10.1109/TIV.2023.3274068

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Intelligent Vehicles

Citation (APA)
Roth, M., & Gavrila, D. M. (2023). IntrApose: Monocular Driver 6 DOF Head Pose Estimation Leveraging Camera Intrinsics. *IEEE Transactions on Intelligent Vehicles*, *8*(8), 4057-4068. https://doi.org/10.1109/TIV.2023.3274068

intrApose: Monocular Driver 6 DOF Head Pose Estimation Leveraging Camera Intrinsics

Markus Roth and Dariu M. Gavrila, *Member, IEEE*

***Abstract*—We present intrApose, a novel method for continuous 6 DOF head pose estimation from a single camera image without prior detection or landmark localization. We argue that using camera intrinsics alongside the intensity information is essential for accurate pose estimation. The proposed head pose estimation framework is crop-aware and scale-aware, i.e., it keeps poses estimated within image cut-outs consistent with the whole image. It employs a continuous, differentiable rotation representation that simplifies the overall architecture compared to existing methods. Our method is validated on *DD-Pose*, a challenging real-world in-vehicle driver observation dataset that offers a broad spectrum of poses and occlusion states from naturalistic driving scenarios. In ablation studies we compare rotation and translation errors of intrinsics-aware and -agnostic methods, continuous and discontinuous rotation representations, and data sampling strategies. Experiments show that leveraging camera intrinsics and a continuous rotation representation (SVDO$^+$) results in a *balanced mean angular error* (BMAE) of 5.8° compared to the intrinsics agnostic baseline with a discontinuous rotation representation (14.8°). Furthermore, training with an unbiased data distribution (most driver measurements are close-to-frontal) improved BMAE on the *hard* subset (extreme orientations and occlusions) from 15.3° to 9.5°.**

***Index Terms*—Head pose estimation, driver observation.**

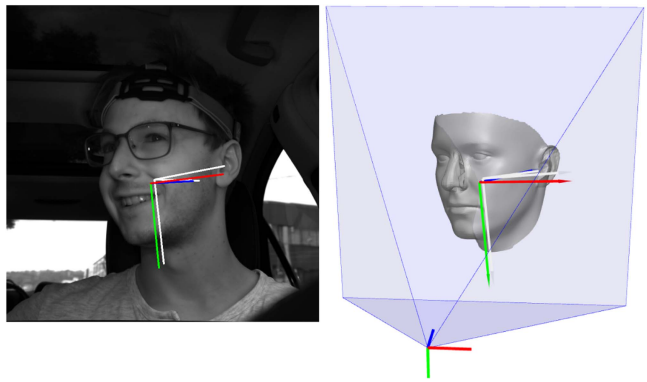

Fig. 1. intrApose is an **intr**inics-**A**ware head **pose** estimation method. The method estimates continuous 6 DOF head pose (rotation and translation) from a single intensity image and known camera intrinsics. Left: Input intensity image (*DD-Pose* validation set). Right: 3D scene with camera frustum (blue). Face mesh and RGB axes: 6 DOF head pose result of intrApose. Gray axes: ground truth head pose.

I. Introduction

HEAD pose estimation plays an essential role in human understanding, as it is our natural cue for inferring focus of attention, awareness, and intention. For machine vision, the task is to estimate both translation and rotation of the head from camera images.

A wide range of uses exist for head pose estimation, either directly or for derived tasks such as gaze estimation, facial identification, facial expression analysis, augmented reality, surveillance, and automotive applications. In the latter, inferring a driver's head pose has been used in safety applications, like estimation of distraction [1], intention, fatigue/drowsiness [2], awareness, and maneuver prediction, e.g., by gaze zone estimation [3], or mirror-checking [4]. It is vital for driver-pedestrian-interaction, e.g., for path-prediction and collision risk estimation [5], or negotiating the right-of-way [6].

Head pose has been employed in on-market vehicles as early as 2007 (Toyota/Lexus) to estimate driver alertness. Cadillac (Super Cruise, 2018), BMW (Extended Traffic Jam Assistant, 2018), and Nissan (ProPilot, 2019) implement extended SAE level 2 capabilities and leverage a driver camera to assess the readiness of the driver to take over the task of driving. Mercedes-Benz's latest S-Class features a stereo driver camera that monitors the driver's readiness to take over from automated driving mode on highways in an SAE level 3 system. This legally allows the driver to perform non-driving related tasks for up to 10 s under specific conditions. In addition, the latest S-Class features a volumetric heads-up display (HUD), an auto-stereoscopic 3D display and multi-modal human-car interaction, each facilitated by head pose.

In-vehicle driver head pose estimation provides particular challenges to vision-based head pose estimation systems due to difficult illumination conditions (such as harsh sunlight covering parts of the face), occlusions (by worn objects such as glasses, but also due to the driving action) and extreme head poses imposed by naturalistic, complex driving scenarios while demanding a precise pose estimate and high availability in a non-invasive setting (no blinding illumination, no worn sensors). These challenges come in addition to the ones vision-based

Manuscript received 4 March 2023; revised 13 April 2023; accepted 23 April 2023. Date of publication 8 May 2023; date of current version 22 September 2023. *(Corresponding author: Markus Roth.)*

Markus Roth is with the Intelligent Vehicles Group at TU Delft, 2628CD Delft, The Netherlands, and also with the Perception and Maps Department at Mercedes-Benz AG, 70546 Stuttgart, Germany (e-mail: markus.r.roth@mercedes-benz.com).

Dariu M. Gavrila is with the Intelligent Vehicles Group at TU Delft, 2628CD Delft, The Netherlands (e-mail: d.m.gavrila@tudelft.nl).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TIV.2023.3274068.

Digital Object Identifier 10.1109/TIV.2023.3274068

human-centric methods have to face, such as the wide variance of appearance, e.g., due to age, gender, ethnicity, or accessories. On the other hand, operating in-vehicle also provides advantages, such as a fixed perspective defined by the known extrinsic and intrinsic camera parameters, and the sparse number of faces simultaneously present within the cabin [7].

Unlike most previous work that have estimated only a subset of the 6 DOF, e.g., by not estimating translation, estimating less than 3 DOF of rotation, or by estimating coarse bins of rotation, we focus on *full* 6 DOF on a *continuous* scale, as required by most of the aforementioned applications. Both the estimation of translation and rotation can be seen as a regression problem, based on the intensity input image. Parameters determining how the head is being projected into the camera image are the intrinsic camera parameters, i.e., focal lengths and principal point. These factors directly affect the estimation accuracy. We will show that a head pose estimation method needs to be *intrinsics-aware* for precise estimation, i.e., being a *camera*-based method rather than an intrinsics-agnostic *image*-based method. On the contrary, a head pose estimation method which does not explicitly consume camera parameters encodes implicit assumptions which hinders generalization to different camera setups.

Stereo cameras bring potential benefits by allowing depth estimates via disparity on a fully calibrated setup, i.e., camera intrinsics for both cameras and the extrinsic rigid transform between them. Yet, a stereo setup brings additional challenges, such as noisy depth estimates, need for (online) re-calibration (e.g., due to thermal deformation of the mounting), synchronization, and a higher cost of hardware.

There are different representations for 3 DOF rotations, most commonly Euler angles and Quaternions. Both come with discontinuities (i.e., Euler angles between 359° and 0°) and non-linearities that we discuss in more detail in Section III-C. We observed that rotation estimation methods have applied workarounds for dealing with the discontinuity of rotation representations rather than intrinsically using a continuous representation [8], [9]. Neural networks are being trained using gradient descent, meaning during error backpropagation a gradient ∇ is being subtracted from the model weights. The adapted weights do not necessarily yield elements of the same space anymore (i.e., $q - \nabla q \notin \mathbb{H}$ for a quaternion $q \in \mathbb{H}$) and needs further post-processing. As a solution, several continuous rotation representations suitable for deep learning methods have been proposed and analyzed in recent years [10], [11]. We follow the definition of Zhou et al. [10]: a rotation representation is *continuous*, if the function f mapping from the representation space R (i.e,. what a neural network estimates) to the original rotation space X (e.g., the special orthogonal group SO(3)) has a continuous function g, such that $f(g(x)) = x \, \forall x \in X$.

Training and evaluating a method that estimates full, continuous 6 DOF head pose demands a dataset that provides camera intrinsics alongside continuous 3 DOF translation annotation and continuous 3 DOF rotation annotation. An additional measurement device or model knowledge about the head geometry is needed for precise ground truth. To that end, we will base our analysis on *DD-Pose* [12], an in-vehicle dataset with precise 6 DOF head pose annotations offering a variety of non-frontal poses and occlusions occurring during the complex, naturalistic driving scenarios.

TABLE I
3 DOF ROTATION REPRESENTATIONS WITHIN DEEP NEURAL NETWORKS AND THE NUMBER OF VALUES (#VAL) THEY ESTIMATE

Representation	#val	Method	⊥	HPE	◌
YPR [8]	3	Euler / Tait–Bryan angles	✓	✓	-
Rotation vector [14]	3	(rotvec). Compact axis-angle	✓	✓	-
Axis-Angle [10]	4	3-vector for axis, scalar angle	-	-	-
Quaternion [15]	4	Quaternion + normalization	-	✓	-
Ortho5D [10]	5	Stereographic projection	✓	-	✓
Ortho6D [10], [16]	6	Gram-Schmidt process	✓	✓	✓
M [10]	9	3x3 matrix (unconstrained)	-	-	-
SVD-inf [9], [11]	9	SVD (inference) on M, ortho loss [9]	-	✓	-
SVDO$^+$ [11]	9	*Differentiable* SVD (training) on M	✓	-	✓

⊥: representation is within SO(3) without post-processing and after each step of backpropagation.
HPE: applied to head pose estimation.
◌: representation is continuous in accordance with the definition of Zhou et al. [10]

In this article, we present intrApose, a method for full, continuous 6 DOF head pose estimation which explicitly leverages camera intrinsics and uses a continuous rotation representation. It operates directly on intensity images and camera intrinsics without a previous face detection or landmark estimation step. See Fig. 1 for exemplary input to and output of our method.

The rest of this article is organized as follows: Section II reviews previous work. Section III defines head pose, analyzes why camera intrinsics matter for accurate head pose estimation and proposes intrApose, a novel 6 DOF head pose estimation method. In Section IV, we evaluate intrApose on *DD-Pose* [12] with respect to rotation and translation accuracy. Section V sets the gained insights into a broader scope. Finally, in Section VI conclusions are drawn.

II. Related Work

In this Section, we review different representations of rotations with a focus on applications within deep neural networks. We further survey methods for image-based head pose estimation.

A. Rotation Representations

There is an abundance of 3 DOF rotation representations, most prominently Euler angles, Tait-Bryan angles, rotation matrices, and quaternions. See Shuster et al. [13] for a survey and Table I for a tabular overview.

Euler angles and Tait-Bryan angles describe a rotation by three rotation components and an implicit or explicit convention of the order of axes the individual rotation components are applied on. In addition, the rotation can be *extrinsic*, defining the rotation about axes of the original coordinate frame which is assumed to be motionless, or *intrinsic*, having the axes rotate along the chain of the three elemental rotations. This results in 12 different conventions for obtaining a well-defined rotation based on three given angles. Euler and Tait-Bryan angle components are restricted on a bound interval, thus offering discontinuities (360° and 0° represent the same amount).

Rotation quaternions ($q \in \mathbb{H}$) are a compact 4-element representation allowing for efficient computation using quaternion algebra. Rotation quaternions suffer from the antipodal problem making q and $-q$ represent the same rotation.

A rotation matrix R ($R \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times3}$, $RR^T = \mathrm{I}$, $\det(R) = +1$) maps an orthonormal basis in $\mathbb{R}^3$ to another orthonormal basis in $\mathbb{R}^3$, spanned by the three columns of R. SO(3) is the special orthogonal group containing all rotations in 3D.

There are less frequently used representations, such as axis-angle (axis $a \in \mathbb{R}^3, ||a||_2 = 1$, angle $\theta \in [0, 2\pi]$), and rotation vectors (rotvec; $r \in \mathbb{R}^3$, with angle $\theta = ||r||_2$).

The above representations have the drawback of being discontinuous, which are less suitable for learning, by leading to higher errors or slower convergence [10]. Recently, rotation representations have been proposed to overcome these drawbacks. Zhou et al. [10] proved that in the three-dimensional space any rotation representation with less than 5 dimensions is discontinuous and thus harder to approximate by a neural network. Zhou et al. [10] construct an Ortho5D and an Ortho6D representation which are both continuous. Out of the 5 (6) values, a rotation matrix $\in$ SO(3) is built using a stereographic projection (a Gram-Schmidt process). Levinson et al. [11] explore the viability of integrating symmetric orthogonalization SVDO$^+$ (based on Singular Value Decomposition (SVD)) directly into the neural network following an unconstrained intermediate representation of 9 values (a degenerate rotation matrix M). SVDO$^+$ is continuous and differentiable, thus suitable within deep neural networks.

B. Head Pose Estimation

Head pose estimation from images has been a popular topic for decades and can be categorized from different perspectives, i.e., the *methodical* perspective, the *I/O (input/output)* perspective, and the *application* perspective.

1) Methodical Perspective: Both surveys of Murphy-Chutorian et al. [17] and Abate et al. [18] use a high-level method-based categorization: template-based methods, subspace-based methods, feature-based methods, and regression-based methods.

Template-based methods estimate head pose by matching appearance templates, i.e. by comparing test images to a set of exemplars with known pose. *Subspace-based methods* map the input space (e.g. image intensities) to a head pose manifold. E.g., Derkach et al. [19] use tensor decomposition to model a non-linear manifold of 3D head poses. *Feature-based methods* make use of an intermediate geometric representation of the face. E.g., Baltrusaitis et al. [20] localize facial landmarks by a constrained local model (CLM) and estimate head pose by a successive generalized adaptive view-based appearance model. Tran et al. [21] fit a 3D morphable model to the head which implicitly encodes head pose. Chang et al. [22], point out several drawbacks of facial landmark locations: they a) are ill-defined, therefore vary in interpretation of the annotator, b) represent facial contours, therefore change with a different viewpoint, and c) become occluded depending on viewpoint. This introduces certain errors in head pose estimates based on facial landmark locations.

Regression-based methods learn a (non-linear) functional mapping from input data to the head pose parameter space. There is an abundance of work within this domain, so we point out some examples representative of the concept and refer the reader to the comprehensive survey on deep regression by Lathuiliere et al. [23]. The methods typically perform regression using a neural network, though other regression models have been applied. Neural network-based regression models consist of a CNN-based backbone for feature extraction and a prediction head. Regressing a discontinuous rotation representation has an impact on the architecture. One prominent scheme is coarse-and-fine/ordinal regression, where coarse bins are classified in addition to continuous regression values [9], [24], [25], [26], [27]. Zhou et al. [8] address the discontinuities of large Euler angles by a wrapped loss. Schwarz et al. [28] choose quaternions and use a regularization term to keep quaternion elements small. Hsu et al. [15] additionally propose losses explicitly dealing with the inter-dependence of certain quaternion elements and the independence of others. Albiero et al. [14] use a rotation vector representation. Their method img2pose estimates the delta from a normalized pose (zero-mean, unit standard deviation) to increase robustness. Further, img2pose employs a calibration point loss which uses a set of head-static 3D points (e.g., 3D head landmarks) and compares the projected points of ground-truth and predicted pose.

Lately, methods have tackled the constraint of orthogonality. Cao et al. [9] propose to estimate the basis vectors or the rotation matrix and use a loss to keep the basis vector close to orthogonal. Yet, SVD is needed to create an orthonormal rotation representation. Zhou et al. [10] have proposed the continuous Ortho6D representation (see Section II-A). Hempel et al. [16] applied it to a head rotation estimation in a simple deep neural network estimating 6 values, and employ the Ortho6D representation. The method does not estimate head translation.

2) I/O (input/output) Perspective: A taxonomy orthogonal to the above structure follows the available input modality (e.g., intensity, depth [29], [30], optical-flow or a combination of those), and whether a single measurement in time is being used or multiple consecutive frames (e.g., tracking [31], [32], RNNs [33]). Another disambiguation is about which of the 3 DOF rotation parameters are being estimated, e.g. from a single yaw angle [34], [35] up to 3 DOF head rotation parameters. Finally, methods can be distinguished by whether they estimate (up to 3 DOF) translation alongside rotation, and whether a preprocessing step is needed before the pose estimation, such as a face bounding box detection. This article focuses on single-image, intensity-based methods estimating continuous, full rotation (3 DOF) and translation (3 DOF).

3) Application Perspective: Based on the application domain, different challenges/requirements arise that need to be addressed by the method. E.g., surveillance applications typically have to deal with low-resolution and tolerate larger rotation errors. A widely used application is head pose estimation within generic images. Generic images are typically easy to obtain (e.g., collected from the internet), but lack other information, such

TABLE II
RELATED METHODS FOR INTENSITY-BASED HEAD POSE ESTIMATION WITH THEIR TRANSLATION AND ROTATION REPRESENTATIONS. CROP-AWARE: WHETHER THE POSE IS CONSISTENT WITH THE IMAGE CROP. IN-VEHICLE: WHETHER THE METHOD HAS BEEN APPLIED TO DRIVER HEAD POSE ESTIMATION. INTR.-AWARE: WHETHER THE METHOD LEVERAGES CAMERA INTRINSICS. NOTE THAT THE TOP FOUR METHODS DO NOT ESTIMATE HEAD TRANSLATION, YET ARE OF INTEREST DUE TO THEIR ROTATION REPRESENTATION

Name	Method	Translation	Rotation Repr.	Crop-aware	In-vehicle	Intr.-aware
TriNet [9]	Coarse-to-fine, SVD-inf	-	R	-	-	-
QuatNet [15]	Quaternion, coarse-and-fine	-	Quaternion	-	-	-
WHENet [8]	Coarse-and-fine, address Euler discontinuities	-	Euler (YPR)	-	-	-
6DRepNet [16]	Ortho6D representation of [10]	-	R	-	-	-
img2pose [14]	Faster R-CNN, crop-invariant proposals	XYZ	Rotation vector	✓	-	-
intraApose (ours)	**Crop-invariant proposals, SVDO+**	**XYZ**	**R**	✓	✓	✓

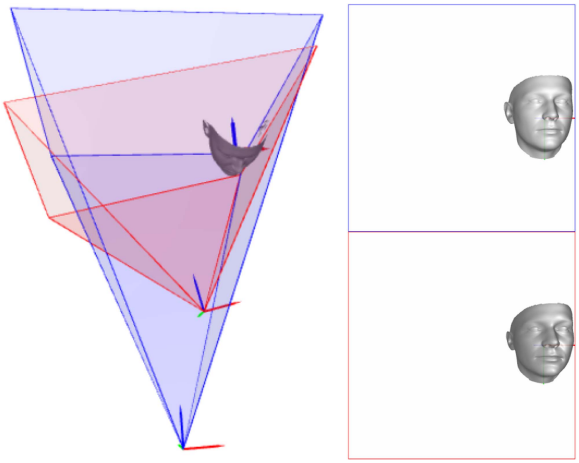

Fig. 2. Motivational example of translation and rotation error introduced by assuming different focal lengths resulting in similar projections. Left: Frusta of two cameras resulting in approximately the same projection of face on the right border of the image. Blue: frustum of camera A with focal length f. Red: frustum of camera B with focal length $f/2$. Camera B is closer to the object due to the larger field of view. In this example, it is rotated against camera A by $> 11°$. The right half depicts the projections into the image space, with camera A on top and camera B with a larger field of view on the bottom.

as camera intrinsics. Within this category one recent method is img2pose [14], a Faster R-CNN-based [36] head pose estimation method which estimates full 6 DOF head pose without prior face or landmark detection. The method regresses bounding boxes out of which features are being pooled for a prediction head. The prediction head regresses a discontinuous rotation vector and a translation vector for each bounding box. The prediction head loses context by being presented with a cut-out of the whole image. Therefore, the bounding-box-local pose is converted to an image-global pose using scaling heuristics. img2pose implicitly assumes a fixed focal length for all input images, leading to erroneous head pose estimates if the assumption fails (i.e., with images depicting a different field-of-view). The ground truth pose used for training and evaluation is obtained using the same focal length assumption and is thus biased.

Another domain is in-vehicle applications. Most approaches focus on head rotation from depth data from structured infrared light, such as Borghi et al. [30], Schwarz et al. [37], or Venturelli et al. [38], while Ahn et al. [39], Firintepe et al. [40] and Schwarz et al. [28] leverage intensity images to estimate head rotation. Out of the aforementioned methods, only Schwarz et al. [37] estimate head translation in addition to rotation, yet only from depth data.

See Table II for the nearest neighbors of the proposed method.

III. INTRAPOSE - 6 DOF DRIVER HEAD POSE ESTIMATION LEVERAGING CAMERA INTRINSICS

A. Overview

We propose intrApose, a novel method for image-based driver head pose estimation based on a deep neural network that regresses continuous 6 DOF from a single intensity image without prior face detection or landmark estimation (see Fig. 3). The main building blocks are a Faster R-CNN-based network which regresses BBoxes and extracts RoI features within these. intrApose learns raw pose features and converts them to a continuous, full 6 DOF head pose within the BBox. This BBox-local pose is converted to an image-global pose in the camera frame while respecting camera intrinsics (see Algorithm 1). Using differentiable modules and a continuous rotation representation allow for a plain overall architecture.

B. Contributions

Our contributions are:

- We observe that neglecting camera intrinsics (e.g., by using heuristics) introduces both rotation and translation errors that exceed reported rotation estimation errors.
  intrApose uses camera intrinsics consistently within the deep neural network and is crop-aware and scale-aware: poses estimated from bounding boxes within the overall image are converted to a consistent pose within the camera frame.
- We borrow for the use in head pose estimation the continuous rotation representation SVDO$^+$ [11] which was used successfully in other domains.
- Using the challenging in-car driver head pose dataset *DD-Pose* [12], we demonstrate that intrApose estimates translation and rotation more robustly compared to state-of-the-art methods, especially for extreme out-of-plane rotations.

Our proposed method is inspired by the recent head pose estimation method img2pose proposed by Albiero et al. [14]. The latter presents an efficient Faster R-CNN-based model

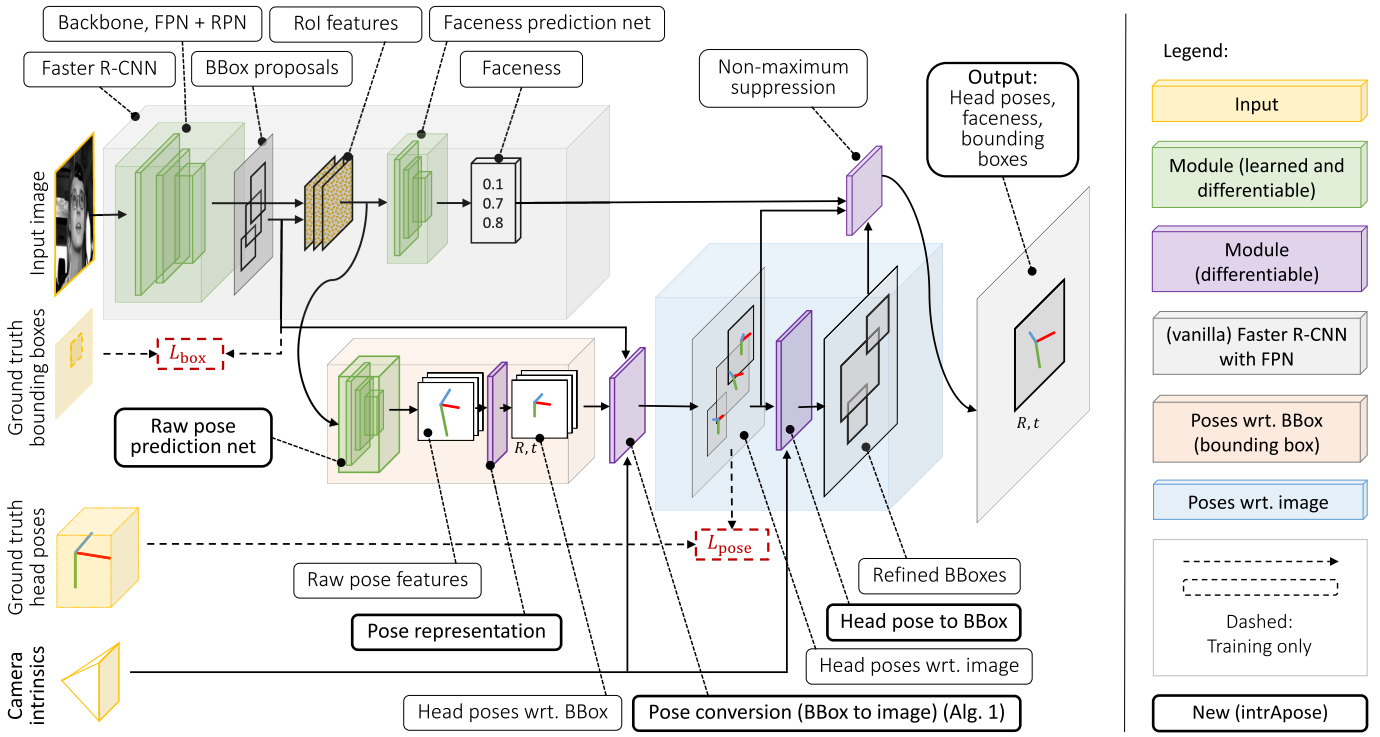

Fig. 3. Architectural overview of intrApose, our proposed head pose estimation method, with novel parts highlighted in bold. intrApose takes intensity images and camera intrinsics as an input. RoI features are obtained from BBox (bounding box) proposals based on Faster R-CNN with feature pyramid network (FPN) [36], [41] (gray box). A **Raw pose prediction net** regresses raw pose features which the **Pose representation** module converts to head poses $\in$ SE(3). Up to here, the poses are relative to their respective BBox (orange box). The BBox-local poses are converted to be image-global (blue box); see Algorithm 1 and Fig. 4. **Bounding boxes are obtained based on the predicted head poses**. A non-maximum suppression step filters overlapping predictions. The output is a set of head poses, faceness scores and bounding boxes. During training, losses are applied to BBox proposals and head poses (dashed lines). The whole architecture is *intrinsics-aware*, specifically in the **Pose conversion (BBox to image)** module and the **Head pose to BBox** projection module, but also with respect to augmentations (see Section III-F) and cropping/resizing.

which regresses 6 DOF head poses without prior face detection or landmark localization. The method has shown strong performance on datasets with ground truth head poses obtained from manually annotated facial landmarks.

The main differences are: a) intrApose is camera-intrinsic aware: focal lengths are consistently used as opposed to using image size as a heuristic for focal length. b) intrApose uses a continuous rotation representation which makes both pose normalization and usage of a calibration point loss as employed by img2pose obsolete, therefore simplifying the architecture. c) intrApose provides an architecture with a differentiable pose conversion which makes an inverse image-to-bbox pose conversion step (i.e., inverse of Algorithm 1) at training time superfluous, therefore further reducing model implementation complexity. d) intrApose uses Faster R-CNN anchor box aspect ratios and sizes tuned for human heads, as opposed to aspect ratios and sizes of generic objects, such as cars or cats.

C. Definition of Head Pose

We define *head pose* as a linear homogeneous transformation matrix $T^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$, the special Euclidean group, which transforms a homogeneous point $p^{\text{head}} = [x^{\text{head}}, y^{\text{head}}, z^{\text{head}}, 1]^T$ given in the *head* coordinate frame to a point $p^{\text{cam}} = [x^{\text{cam}}, y^{\text{cam}}, z^{\text{cam}}, 1]^T$ in the *cam* coordinate frame by $p^{\text{cam}} = T^{\text{cam}\leftarrow\text{head}} \cdot p^{\text{head}}$, thus representing translation by 3 DOF and rotation by 3 DOF on a continuous scale.

Transforms $T \in \text{SE}(3)$ are constructed as in (1). They can be decomposed into a 3×3 submatrix $R \in \text{SO}(3)$ representing the rotation and a translation vector $t = [t_x, t_y, t_z]$. Ultimately, a homogeneous point multiplied from the right-hand side will be rotated by R and afterward shifted by t.

$$T^{\text{cam}\leftarrow\text{head}} = \begin{bmatrix} & R & & \begin{matrix} t_x \\ t_y \\ t_z \end{matrix} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

For the *cam* frame, we follow the convention: x to the right, y to the bottom, and z in the viewing direction. The *head* frame can be an arbitrary, head static frame.

D. Why Camera Intrinsics are Essential for Pose Estimation

The camera intrinsic matrix K defines, how 3D points in the *cam* frame are projected onto a rectified image. See (2):

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

It is a 3×3 matrix consisting of focal lengths f_x and f_y for x and y axes and principal point (c_x, c_y) representing the optical center

within the image. The axis skew parameter s is typically assumed 0. K projects a point $p^{\text{cam}} = [x^{\text{cam}}, y^{\text{cam}}, z^{\text{cam}}]^T$ given in the *cam* frame onto pixel coordinates $[u, v]$ by $[u \cdot w, v \cdot w, w]^T = K \cdot p^{\text{cam}}$. Points residing in a different frame, e.g., *head*, can be transformed into the *cam* frame by a rigid transform $T^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$. K is non-singular: its inverse K^{-1} projects an image coordinate $[u, v, 1]$ into a 3D ray representing all points in *cam* frame which project onto $[u, v, 1]$.

Translation and Rotation Errors: One implication of *assuming* focal lengths not corresponding to the camera optics, e.g., kf_x and kf_y for a factor $k \in \mathbb{R}$ will project a central object to k times the image size (k^{-1} times the distance to the camera), compared to focal lengths f_x and f_y. When estimating object pose, its z translation will be k times as large.

Another implication is that assuming a wrong camera intrinsic matrix affects rotation estimations which are more apparent at the image border. Take the example in Fig. 2: two cameras differing in focal lengths by a factor of two are positioned such that their projections of a head pose into the respective camera images are approximately equal (close to the right image border). The camera with a larger field of view (smaller focal length) is closer to the head and rotated. A pose-from-image estimation using these different camera intrinsics from the same image results in a translation error of factor two and a rotation error of $> 11°$ (in this example).[1]

E. Proposed Model

See Fig. 3 for an architectural overview. Given an image $\mathbf{I}$, we estimate full 6 DOF continuous head pose $T_i^{\text{cam}\leftarrow\text{head}}$ for each head i within the image $\mathbf{I}$. The major building blocks are a Faster R-CNN module which predicts bounding box proposals along with a faceness score. Our prediction head performs RoI pooling on the backbone's feature maps based on the bounding box proposals to obtain RoI features. A *raw pose prediction net* predicts an intermediate, unconstrained representation of *raw pose features* which are typically of small size (such as 6 to 12 values, see Table I) representing rotation and translation. The *Pose representation* converts the potentially degenerate raw pose features to a head pose $\in \text{SE}(3)$ wrt. the isolated BBox. A *Pose conversion* module converts BBox-local head pose to an image-global pose such that it projects approximately equal within the (whole) image. This essentially performs scaling and rotation of the pose based on image intrinsics and bounding box location and size (see Algorithm 1 and Fig. 4). The final step is a non-maximum suppression based on projected bounding boxes and faceness scores and yields a set of head poses. Let us describe the components in more detail.

Backbone, FPN + RPN: intrApose extends the two-stage object detection approach of Faster R-CNN [36] with Feature Pyramid Networks (FPN) [41] by a head pose estimation module. Faster R-CNN consists of a backbone network that extracts features on multiple scales from the input image. Using these features and anchor bounding boxes of typical object aspect ratio and shape, a region proposal network (RPN) predicts bounding box proposals alongside an objectness score. An RoI pooling operation aggregates features for each bounding box proposal into *RoI features*.

[1]The estimated poses using camera A and camera B would differ by the rigid transform which brings camera A into camera B.

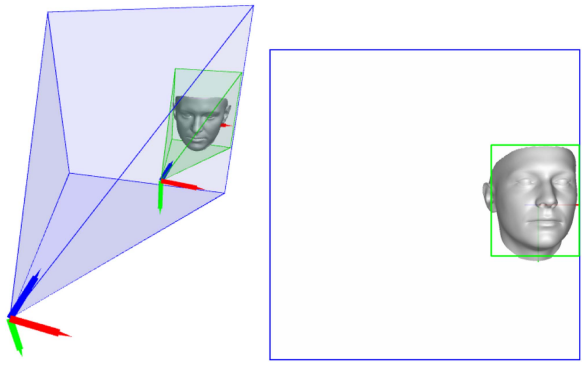

Fig. 4. Illustration of the pose conversion of Algorithm 1. Left: Frusta of whole image camera C_{image} (blue), BBox camera C_{bbox} (green) and a head pose (mesh). Right: projection of the 3D scene into the camera image using K_{image} and a bounding box of the head (green). C_{bbox} is a virtual camera with a focal length proportional to its bounding box size, thus representing a canonical size. C_{bbox} is closer to the head pose and the principal axis goes through the bounding box center. As a result, a nearly frontal pose estimated within the bounding box will be converted to a head pose close to the right border of the whole image, rotated and further away from C_{image}.

Head pose estimation module: As in img2pose [14], we propose a network that regresses a BBox-local head pose and a faceness score p_i for each RoI feature map i. In contrast to img2pose, which regresses 6-element vectors representing head pose directly (rotvec and translation), our architecture allows for a generic scheme by estimating an intermediate raw pose feature representation by a *Raw pose prediction net* that is being converted to a head pose $T_i^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$ by a differentiable *Pose Representation* module.

Raw pose prediction net: The raw pose prediction net estimates unconstrained, raw pose features $f_{\text{R,t}}$ (R: rotation, t: translation) wrt. the BBox for each RoI feature map. It consists of a batch-normalized fully connected layer with 256 features and ReLU activation followed by another fully connected layer reducing to the number of raw pose features $f_{\text{R,t}}$.

Pose representation: The *Pose representation* module converts the raw pose features into a $T_{\text{bbox}}^{\text{cam}\leftarrow\text{head}} \in \text{SE}(3)$ representation. We have seen in Section II that there is a number of pose representations available. As Zhou et al. [10] and Levinson et al. [11] pointed out, a continuous, differentiable rotation representation is favorable. We will analyze different rotation representations, such as the (discontinuous) rotation vector representation of Albiero et al. [14], but also Ortho6D and SVDO$^+$. The translation part of the pose is treated in a regular manner, meaning 3 DOF metric translation is being regressed.

From an integration perspective, both Ortho6D and SVDO$^+$ take 6, respectively 9 unconstrained values as an input and create a rotation matrix $R \in \text{SO}(3)$. One important aspect is that the pose representation needs to be differentiable to allow for gradients to pass during training. Both Ortho6D and SVDO$^+$ are differentiable.

Algorithm 1: Pose Conversion (BBox to Image).

```
def convert_pose_bbox_to_image(T_cam_head, bbox, K_image):
  # create bbox intrinsic matrix with same focal lengths
  # as image and principal point in center of bbox
  K_bbox = copy(K_image)
  K_bbox(cx) = get_center_u(bbox)
  K_bbox(cy) = get_center_v(bbox)

  # scale: ratio of image focal length and
  # bbox size (canonical bbox camera)
  f_image = K_image(fx)
  size_bbox = get_w(bbox) + get_h(bbox)
  scale = f_image / size_bbox
  T_cam_head(z) *= scale # scale z

  # apply 4x4 homography matrix to head pose
  H_image_bbox = homogen(inv(K_image) @ K_bbox)
  T_cam_head = H_image_bbox @ T_cam_head
  T_cam_head = orthogonalize_svdo+(T_cam_head)

  return T_cam_head # image-global
```

Note that using a continuous pose representation allows designing a network without bells and whistles, i.e. no coarse-to-fine approach, no estimation of delta to a mean pose, no dealing with values at discontinuities, etc.

Pose conversion (BBox to image): As posed by Albiero et al. [14], each BBox proposal is a cut-out of the image and lost its information about the location within the image. Therefore, head poses are estimated wrt. their BBox and need conversion to the full image. To that end, we propose an intrinsics-, crop- and scale-aware BBox pose to image pose conversion method that extends the conversion method of img2pose [14] and is described in Algorithm 1 and illustrated in Fig. 4. In essence, the conversion method builds a homogeneous canonical BBox camera matrix K_{bbox} which has the same focal length as the image camera matrix K_{image}, and the principal point in the bounding box center. The distance-to-camera t_z is being scaled by the ratio of image focal length to bounding box size. Therefore, BBox head pose is estimated within a canonical BBox camera. Scaling accounts for the fact that a cut-out with a close-by head is tightly enclosed by the bounding box and reflects a head further away in the image. Inter-individual head sizes are learned implicitly from the training data. The pose is afterward projected into the pixel space with K_{bbox} and back into the 3D space of the camera with the inverse of K_{image}. The homography $K_{\text{image}}^{-1} K_{\text{bbox}}$ is not orthonormal, meaning it does not keep the basis vectors of the transform orthogonal and in unit length. Therefore, a successive (differentiable) orthogonalization of the rotation is necessary to stay within SE(3).[2] Overall, Algorithm 1 makes the method intrinsics-, crop- and scale-aware.

Head Pose to Box: With head pose and camera intrinsics, we can get well-defined bounding boxes at minimum additional cost. If we define bounding boxes as a rectangle in the image which encloses all parts of the object of interest, then we can define 3D points representing extrema in the head frame (chin to the top of forehead, nose, ears), transform them into the camera frame and project them into image space using K_{image}. The image bounding box is defined by the extrema of the projected pixel coordinates. A margin can be defined either in 3D space (by taking 3D points outside a typical head), or in image space (by adding a margin to the projected bounding box).

Training objective: During training, we aim to optimize the following objectives: a) RPN bounding box proposals (vanilla Faster R-CNN), b) RPN objectness score (vanilla Faster R-CNN), c) Faceness score, and d) Head pose outputs wrt. image, which is a multi-task problem. For a) and b) we refer to Ren et al. [36] (L_{bbox}: smooth L_1 on positive samples; $L_{\text{objectness}}$: binary cross entropy L_{cls}). For the other representations, we define the following loss functions.

Faceness score: We match ground truth bounding boxes (automatically generated from head pose ground truth) with proposal bounding boxes using Intersection-over-Union (IoU). Positive matches yield a faceness loss of $L_{\text{face}} = L_{\text{cls}}(p_i, 1)$ for the predicted faceness score p_i. Negative matches get $L_{\text{face}} = L_{\text{cls}}(p_i, 0)$.

Head pose: The head pose matrix $T^{\text{cam}\leftarrow\text{head}}$ can be decomposed into a 3×3 rotation matrix R and a translation vector $t = [t_x, t_y, t_z]^T$ as in (1). Positive matches are considered for a head pose loss $L_{\text{pose}} = L_{\text{R}} + L_{\text{t}}$, consisting of rotation loss L_{R} and translation loss L_{t}. We apply the loss to predicted poses wrt. full image.

We define the translation loss $L_{\text{t}}(t, \hat{t}) = ||t - \hat{t}||_2^2$ for the estimated translation t and the ground truth translation $\hat{t}$. The rotation loss $L_{\text{R}}(R, \hat{R}) = L_{\text{geodesic}} = \arccos\left(\frac{\text{tr}(R\hat{R}^T)-1}{2}\right)$ corresponds to the geodesic distance between the predicted rotation R and the ground truth $\hat{R}$. We optimize the overall loss L:

$$L = L_{\text{objectness}} + L_{\text{bbox}} + L_{\text{faceness}} + L_{\text{R}} + L_{\text{t}} \tag{3}$$

F. Intrinsics-Consistent Image and Pose Augmentations

Augmentations are a scheme to create further training data to obtain a more robust model. The ground truth to our model is given by the tuple (image (h, w), camera intrinsics K, head pose $T^{\text{cam}\leftarrow\text{head}}$). The invariant of each augmentation is that the tuple remains consistent in the sense that the augmented head pose is being projected onto the corresponding locations of the augmented image using the augmented camera intrinsics. We implemented intrinsics-aware crop, scale, and flip augmentations.

Crop image with $bbox_{\text{crop}} = [u, v, w, h]$ needs a shift of the principal point for the augmented camera intrinsics: $K_{\text{crop}} = \begin{bmatrix} f_x & 0 & c_x - u \\ 0 & f_y & c_y - v \\ 0 & 0 & 1 \end{bmatrix}$. Head pose remains the same.

Scaling image height/width with factor s_h/s_w needs rescaling of focal lengths and principal point: $K_{\text{scale}} = \text{diag}(s_w, s_h, 1) \cdot K$.

Flip (left-right) flips the principal point c_x: $K_{\text{flip}} = \begin{bmatrix} f_x & 0 & w - c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$. The head pose needs to be flipped on the yz-plane of the camera frame, which can be obtained by the

[2]Note that Albiero et al. [14] do not explicitly formalize the orthogonalization of the degenerate rotation. In their reference implementation, orthogonalization happens implicitly during the pose conversion step from rotation matrix to rotation vector (`rot_mat_to_rot_vec()`). We explicitly formalize this step.

following Hadamard product ($\circ$, piecewise multiplication) and keeps the transform right-handed:

$$T^{\text{cam}\leftarrow\text{head}}_{\text{flipped}} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \circ T^{\text{cam}\leftarrow\text{head}}$$

G. Training Details

We implemented the proposed intrApose model in PyTorch with a ResNet-18 backbone [42] which was pretrained on natural images. All implemented modules are differentiable to allow gradients to flow backward from the losses. This includes orthogonalization and pose conversion modules. We used stochastic gradient descent (SGD) on mini-batches of four images with an initial learning rate of 0.001 and a weight decay of $5 \cdot 10^{-4}$. We reduce the learning rate by a factor of 10 if the model has not improved over the last three epochs on the validation set. Similarly, we perform early stopping after 5 epochs without improvement on the validation set.

For training the RPN, we randomly sampled 256 bounding box proposals per image. For training the faceness prediction net and the raw pose prediction net, we sampled 512 proposals per image.

We augmented the training data by intrinsics-aware scaling, mirroring, and cropping, as detailed in Section III-F. Unbiasing: To make the model more robust in non-frontal poses, training samples with non-frontal poses are sampled more frequently compared to the dataset distribution which typically consists of more frontal driver head poses in in-car settings.

Training converged after 11 epochs and took approximately 2.5 days on a single NVidia Tesla V100 GPU. The model has $4.2 \cdot 10^7$ parameters. Inference time on the `float32` model is 18.4 samples per second.

IV. Experiments

A. Dataset

Evaluation of intrApose puts specific requirements on the evaluation dataset. Datasets that do not provide camera intrinsics become out of scope as argued in Section III-D. We therefore chose *DD-Pose* [12], a large-scale in-car dataset (330 k images) that offers complex naturalistic driving scenarios. 6 DOF continuous head pose annotations are provided with the help of an optical marker tracker which is worn by the driver. The head-worn marker is spatially calibrated once for each of the 27 drivers and results in head pose measurements free of drift and latency. The complex driving scenarios offer a broad distribution of head poses (see Fig. 5), yet representative of typical driving, including challenging lighting conditions. *DD-Pose* offers dataset splits depending on occlusion annotation and angle-from-frontal (all, easy, moderate, hard). We chose subjects {8, 19, 23} for validation, subjects {3, 6, 10, 11, 14, 15, 16, 17} for testing (as defined by Roth et al. [12]) and the other subjects for training.

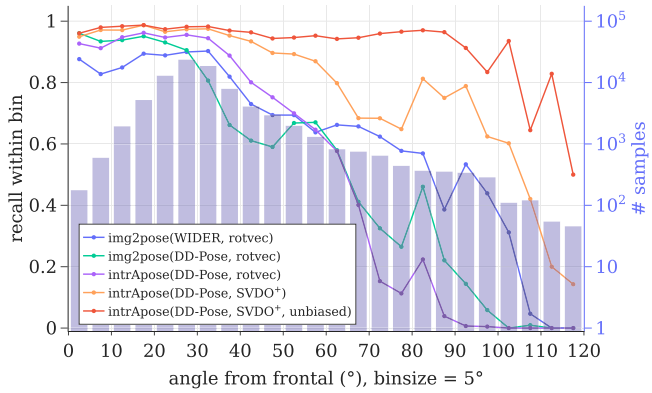

Fig. 5. Recall and data distribution over the angular difference from frontal pose for the *DD-Pose* test set. Compared to the baseline (*img2pose(WIDER, rotvec)*), the recall improves incrementally by training on *DD-Pose* (*img2pose(DD-Pose, rotvec)*), switching to the proposed architecture (*intrApose(DD-Pose, rotvec)*), using a continuous rotation representation (*intrApose(DD-Pose, SVDO$^+$)*) and training using an unbiased dataset with more non-frontal poses (*intrApose(DD-Pose, SVDO$^+$, unbiased)*).

B. Model Variants

We conduct our experiments with 5 different approaches, distinguished by model used (img2pose vs. intrApose), training dataset (WIDER vs. *DD-Pose*), and rotational representation (rotvec vs. SVDO$^+$). We evaluate recall, translation error, and rotation error.

img2pose(WIDER, rotvec): Pretrained img2pose model provided by Albiero et al. [14] (see Section II). The authors trained the model on the WIDER face dataset, which does not provide camera intrinsics. 3D head poses on the WIDER dataset were created by the authors using Perspective-n-Point on facial landmarks with a large head model and an assumed focal length equaling the sum of image width and image height, i.e., the same focal length assumption the method makes internally. One important fact to mention is that using the same focal length for the creation of the ground truth pose imposes a bias. The large head model (width $\approx 1.5\,\text{m}$, 10 times as large as a mean head) introduces head translations about 10 times as far away from the camera.

img2pose(DD-Pose, rotvec): This is the same model as *img2pose(WIDER, rotvec)*, but trained on *DD-Pose* using the training scheme of Albiero et al. [14], i.e., using assumed focal lengths instead of the true camera intrinsics provided with the *DD-Pose* dataset. Bounding boxes and head poses were used as provided by *DD-Pose*. We needed to use 3D head landmarks of typical size (width $\approx 0.15\,\text{m}$) in the calibration point loss of img2pose for the training to converge, potentially caused by points of the large model being projected outside the image in the calibration point loss. Note that the points used by calibration point loss are not to obtain a scale (as with landmark-based approaches), but rather to guide the model in adapting its parameters for pose estimation during training.

intrApose(DD-Pose, rotvec): Our proposed model with the discontinuous rotation vector (rotvec) representation and the L_2 loss function of img2pose. This model is intrinsics-aware. We use pose normalization as in img2pose, i.e., estimating the

TABLE III
ROTATION ERRORS, TRANSLATION ERRORS AND RECALL ON THE *DD-POSE* TEST SET FOR THE MODEL VARIANTS ON DIFFERENT SUBSETS (ALL, E: EASY, M: MODERATE, H: HARD). SEE SECTION IV-B FOR DETAILS ON THE MODELS. ROTATION ERRORS ARE GIVEN IN DEGREES (°), AND TRANSLATION ERRORS IN MILLIMETERS (mm)

Method	BMAE (°) ↓				MAE_R (°) ↓				MAE_t (mm) ↓				recall (%) ↑			
	all	e	m	h	all	e	m	h	all	e	m	h	all	e	m	h
img2pose(WIDER, rotvec) [14]	10.3	6.4	11.1	20.3	7.8	6.7	9.4	18.4	7849	7746	8068	8431	85	99	64	56
img2pose(DD-Pose, rotvec)	14.8	5.9	12.5	48.3	6.9	5.1	8.7	42.6	**18**	**14**	23	78	81	92	68	33
intrApose(DD-Pose, rotvec)	7.5	6.4	7.5	12.6	6.3	6.0	6.8	9.7	19	18	**21**	**26**	89	99	80	24
intrApose(DD-Pose, SVDO$^+$)	8.0	**4.0**	8.0	15.3	5.0	4.0	**5.9**	16.0	21	18	24	47	95	**100**	90	71
intrApose(DD-Pose, SVDO$^+$, unbiased)	**5.8**	4.2	**6.2**	**9.5**	**4.8**	**3.9**	**5.9**	**8.9**	25	22	29	41	**97**	**100**	**93**	**93**

↑/↓: higher/lower values denote better performance.

pose with zero-centered mean and unit standard deviation. We use a head model of typical size for the calibration point loss and applied the proposed intrinsics-aware crop, flip, and scale augmentations defined in Section III-F.

intrApose(DD-Pose, SVDO$^+$): Our proposed model with the continuous pose representation SVDO$^+$ [11] and geodesic loss. The model is intrinsics-aware. We found pose normalization to be unnecessary. We tuned anchor sizes and aspect ratios on the *DD-Pose* training set. Compared to img2pose, no point calibration loss was necessary. We used a typical scale head model to create bounding boxes from the predicted head poses.

intrApose(DD-Pose, SVDO$^+$, unbiased): Same model as *intrApose(DD-Pose, SVDO$^+$)*, but trained with an unbiased dataset by sampling more non-frontal poses.

C. Recall

Recall defines on which percentage of the images a head hypothesis from head pose estimation method exists. Images without a hypothesis are left out when evaluating translation and rotation. For matching ground truth and hypotheses, we use an Intersection-over-Union (IoU) threshold of 0.3 for the respective bounding boxes. Predicted head poses with a faceness score of > 0.9 are considered.

Fig. 5 depicts the recall over the angle difference from frontal head pose. The rotation vector-based methods (rotvec) have a recall of > 0.8 for frontal faces and drop towards 0.6 for rotations $60°$ off-frontal. Out of these, the model variants trained on *DD-Pose* have a higher recall for close-to-frontal poses. The baseline *img2pose(WIDER, rotvec)* offers a higher recall for highly off-frontal faces ($[70°, 100°]$) compared to the other rotation vector-based methods. We explain this by the WIDER dataset having a more homogeneous angular distribution compared to *DD-Pose*, which offers more close-to-frontal faces (see histogram in Fig. 5).

Using the continuous SVDO$^+$ rotation representation (*intrApose(DD-Pose, SVDO$^+$)*) shows a considerable benefit across the whole angular spectrum compared to the rotation vector representation (*intrApose(DD-Pose, rotvec)*), keeping the recall above 0.6 for angles up to $105°$ and dropping towards 0.4 for $110°$. The same model trained with an unbiased dataset (*intrApose(DD-Pose, SVDO$^+$, unbiased)* shows remarkable improvement of recall for extreme poses, keeping the recall above 0.8 across the whole angular spectrum until $105°$, only afterward dropping towards 0.4. The right side of Table III shows recall aggregated over the subsets (all, easy, moderate, hard) in accordance with the observations from Fig. 5.

D. Translation Error

We evaluate the mean Euclidean distance $\mathrm{MAE_t}$ between ground truth head origin and predicted head origin.

The errors in head translation estimation ($\mathrm{MAE_t}$) are listed in Table III. The pretrained baseline *img2pose(WIDER, rotvec)* depicts an error of over $7.7\,\mathrm{m}$. Overestimated distance to camera (t_z) contributes most to the error. This is caused by two facts: for one, *img2pose(WIDER, rotvec)* assumes a focal length defined by the image size which does not correspond to the true intrinsics of the camera. Also, the WIDER dataset consists of 2D facial landmark labels which Albiero et al. use to generate the ground truth head poses by Perspective-n-Point and a 3D head model which is $\sim 1.5\,\mathrm{m}$ wide. The model trained on WIDER therefore estimates heads presented in *DD-Pose* further away. As Albiero et al. use WIDER for both training and evaluation, this fact had not become apparent. In comparison, the img2pose-based model trained on *DD-Pose* (*img2pose(DD-Pose, rotvec)*) shows a better estimation of the head translation, caused by the correct head pose ground truth obtained by a measurement device. Overall, the head is estimated $18\,\mathrm{mm}$ from the ground truth for the *all* subset and $78\,\mathrm{mm}$ for the *hard* subset. The non-unbiased intrApose model variants perform similarly in translation estimation, being less than $21\,\mathrm{mm}$ off for the *all* subset. Comparing the SVDO$^+$ model variants shows that unbiasing the training dataset decreases translation error from $47\,\mathrm{mm}$ to $41\,\mathrm{mm}$ on the *hard* subset, though sacrificing $\mathrm{MAE_t}$ for the other subsets (*easy*, *moderate*). We explain the worsening on the latter subsets by the use of significantly fewer training samples from these subsets while evaluation is biased in the sense that a large portion of samples resides in the *easy* and *moderate* subsets (see data distribution in Fig. 5). Another hypothesis is a higher imbalance of rotation loss and translation loss (unbiasing is based on angle from frontal).

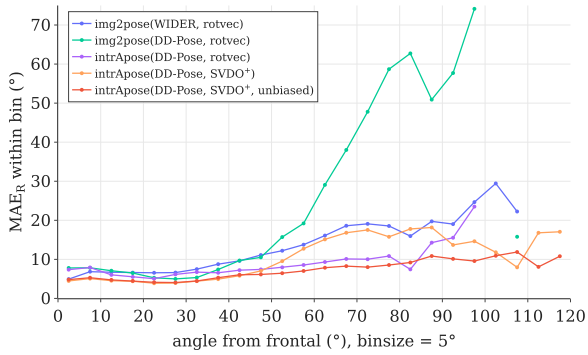

Fig. 6. Mean angular errors ($\mathrm{MAE_R}$) on the *DD-Pose* test set. $\mathrm{MAE_R}$ increases with larger angular distance from frontal. The model variant *intrApose(DD-Pose, SVDO$^+$, unbiased)* provides a consistently low angular error over the whole depicted angular spectrum.

E. Rotation Error

We evaluate rotation error by *mean angular error* $\mathrm{MAE_R}$ of the geodesic distance between ground truth rotation and predicted rotation. For an unbiased evaluation of head rotation, we use *balanced mean angular error* (BMAE) as proposed by Schwarz et al. [28]. It splits the dataset in bins based on the geodesic distance from the frontal pose and averages the $\mathrm{MAE_R}$ of the bins:

$$\mathrm{BMAE}_{d,k} := \frac{d}{k} \sum_i \phi_{i,i+d}, i \in d\mathbb{N} \cap [0, k]$$

where $\phi_{i,i+d}$ is the $\mathrm{MAE_R}$ of all hypotheses, where the geodesic distance between ground truth and frontal pose is between i and $i + d$. During evaluation, we use bin size $d := 5°$ and maximum angle $k := 120°$.

The overall mean angular error ($\mathrm{MAE_R}$) and balanced mean angular error (BMAE) are displayed in Table III and the $\mathrm{MAE_R}$ over the angular difference from a frontal pose are depicted in Fig. 6.

The pretrained baseline *img2pose(WIDER, rotvec)* shows a $\mathrm{MAE_R}$/BMAE of 7.8°/10.3° on the *all* subset of *DD-Pose*, though being trained on WIDER, a dataset based on images downloaded from the internet, therefore shows good generalization to unseen data.

Retraining the img2pose model on *DD-Pose* (*img2pose(DD-Pose, rotvec)*) decreased the $\mathrm{MAE_R}$ to 6.9°, yet increasing the BMAE to 14.8°. This is due to the worse performance for non-frontal poses (see increasing $\mathrm{MAE_R}$ with increasing angle-from-frontal in Fig. 6). We explain this by the majority of the training samples within *DD-Pose* being close-to-frontal, making the model tend to estimate the mean pose with the discontinuous rotation vector representation.

intrApose(DD-Pose , rotvec) uses the same data and pose representation within the proposed intrinsics-aware intrApose framework including the proposed augmentations. We can see the BMAE decrease to 7.5° on the *all* subset and considerably improve on the *hard* subset to 12.6° (from 20.3° and 48.3° of the img2pose models trained on WIDER and *DD-Pose*, respectively). We attribute this improvement to the intrinsic-aware model.

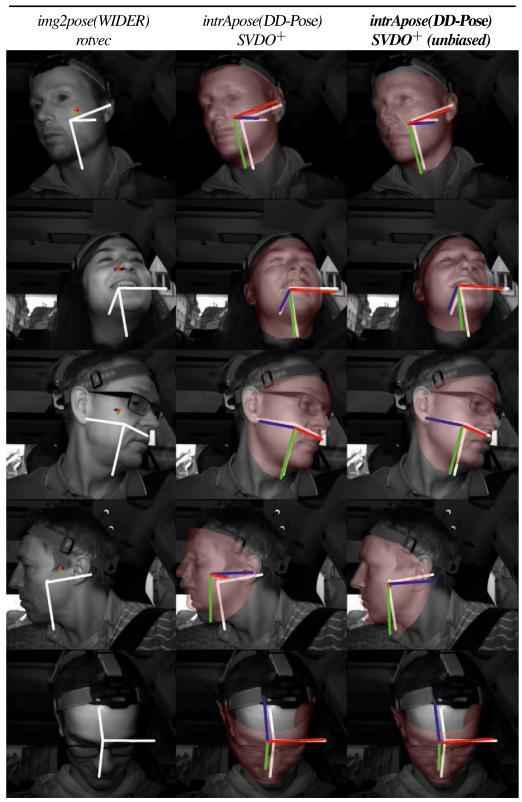

Fig. 7. Qualitative head pose estimation results on samples with challenging off-frontal head poses. We project the poses into the camera image using the camera intrinsics. Ground truth head pose is denoted by a white axis. Predicted head pose is denoted with an RGB axis and a transparent red face mesh of typical head size. The translation error can be judged by comparing the axis length. All images are crops to ease judgement.

Switching to the continuous rotation representation SVDO$^+$ (*intrApose(DD-Pose, SVDO$^+$)*) decreases the $\mathrm{MAE_R}$ to 5.0°, yet increases in terms of BMAE (8.0°). A look at the corresponding recall on the *hard* subset shows that it now predicts more extreme poses (71% vs. 24%) and still tunes towards close-to-frontal poses, as shown by the best $\mathrm{MAE_R}$ on the *easy* subset. Overall, one can say that the closer the BMAE of a model is to the $\mathrm{MAE_R}$, the better it covers data-imbalance.

The final, proposed model *intrApose(DD-Pose, SVDO$^+$, unbiased)* resolves this data imbalance by being trained with more off-frontal pose samples. We can see an improvement of both $\mathrm{MAE_R}$ and BMAE on the hard subset of *DD-Pose*. The model variant shows a consistently low error along the full spectrum of angles from frontal (Fig. 6) and results in a BMAE of 9.5° on the *hard* subset, being very close to the corresponding MAE of 8.9°.

Experiments with the continuous Ortho6D rotation representation of Zhou et al. [10] showed results similar to the SVDO$^+$ rotation representation, in accordance with the observations of Levinson et al. [11].

F. Qualitative Results

Fig. 7 provides qualitative results of the baseline img2pose [14] and the proposed model. The small axes of

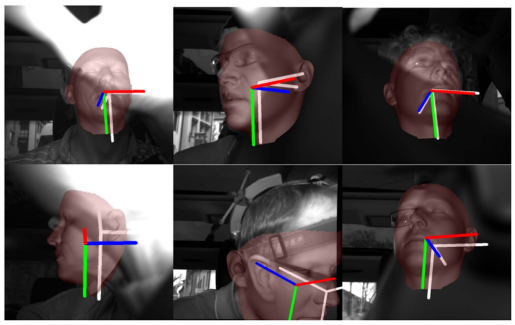

Fig. 8. Random subset of samples of the *DD-Pose* test set where a head pose estimate could only be provided by our proposed model (unbiased).

img2pose confirm the overestimated head translation observed in Table III. As designed, the unbiased proposed model depicts a smaller qualitative error for off-frontal poses compared to the unbiased variant. Samples where only the proposed model could provide a head pose estimate (faceness > 0.9) are depicted in Fig. 8. The model shows robustness towards high occlusions by hands and steering wheel, and extreme poses, though with a larger qualitative error compared to the samples given in Fig. 7.

V. Discussion

We presented a 6 DOF head pose estimation method which employs a continuous rotation representation. For more than two decades, authors have committed to Euler angles or quaternions and treated the values as a simple regression problem, thus ignoring the underlying manifold at hand. This led to complex mitigations dealing with the drawbacks of the representations, such as coarse-to-fine approaches, normalizing values (zero mean, unit standard deviation, quaternion normalization), explicitly handling discontinuities (e.g., at 360°) or proposing special losses (e.g., by encouraging orthogonality or projection of calibration points). This article confirmed the importance of the proper choice of rotation representation of Levinson et al. [11]: we could represent 3 DOF rotation without special pre- or postprocessing, thus leading to a plain network without bells and whistles.

When evaluating, representing rotation errors by Euler angles shows drawbacks due to their ambiguity (order of axes, direction, handedness). Therefore, we reside to geodesic distance, making it agnostic of single angle components and the frame, but at the cost of lacking insight into the contribution of individual axes to the geodesic distance.

Our architecture is based on the Faster R-CNN framework. In general, it can be adapted to other, potentially deeper backbones or to single-shot detection networks.

Our method is intrinsics-aware, therefore requiring camera parameters alongside the image itself. In Section III-D we showed that assuming incorrect camera intrinsics can introduce large errors beyond accepted tolerances. With missing calibration information the error behavior of the proposed method assimilates to methods that are intrinsics-agnostic.

The pose prediction network operates on feature RoIs, therefore can only estimate a local pose within the BBox. Albiero et al. [14] proposed a pose conversion to the whole image. We generalized the pose conversion algorithm by making it intrinsics-aware, allowing for generic pose representations, and explicitly formalizing a necessary orthogonalization step. Essentially, this shows that the pose conversion is a rigid coordinate transformation that approximates the projections of the BBox and the whole image.

Rendering a face mesh overlay is an appealing visualization of head pose. However, using a fully opaque one makes the viewer tolerate more errors both in rotation and translation, by still appearing natural. To that end, we suggest rendering the face mesh transparent and also visualizing the frame axes.

The evaluations have shown that the proposed model provides robust per-frame pose estimates, also for large out-of-plane rotations. However, a recall of 93% on the *hard* subset might be insufficient for safety-relevant in-vehicle applications. A driver-monitoring system would integrate these 6 DOF measurements in a temporal filtering scheme to increase the recall. Such a system could also benefit from *estimated* uncertainties for all 6 DOF. To that end, Bingham belief [43] could be a fitting representation for rotation uncertainties in SO(3).

VI. Conclusion

This manuscript has tackled the problem of 6 DOF head pose estimation from images and their associated camera intrinsics in the domain of driver-observation. This domain poses interesting in-car applications and challenges such as difficult illumination conditions and large out-of-plane rotations.

We showed that explicit use of camera intrinsics is required for precise head pose estimation and use it consistently within our novel intrinsics-aware head pose estimation method.

Discontinuous rotation representations such as Euler angles and quaternions have shown drawbacks that led to complex architectures. Our method employs a continuous rotation representation (SVDO^+) which simplifies the network architecture to a simple regression head and a pose conversion which yields a rotation in SO(3).

Evaluations on the challenging in-car dataset *DD-Pose* have shown that leveraging camera-intrinsics alongside a continuous rotation representation results in a *balanced mean angular error* (BMAE) of 5.8° compared to the intrinsics-agnostic baseline (14.8°). Also, using an unbiasing data sampling strategy lowered the BMAE on the *hard* subset (extreme rotations and occlusions) from 15.3° to 9.5°. The proposed method showed translation errors of $22/29/41$ mm over the *easy*/*moderate*/*hard* subsets in the *DD-Pose* test set.

Overall, our proposed method allowed for a simple architecture that yields robust head pose estimates across a broad spectrum of head poses. Future work involves integration of uncertainties, e.g., by Bingham belief, as well as state estimation over consecutive timesteps.

References

[1] A. El Khatib, C. Ou, and F. Karray, "Driver inattention detection in the context of next-generation autonomous vehicles design: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4483–4496, Nov. 2020.

[2] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2339–2352, Jun. 2019.

[3] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 254–265, Sep. 2018.
[4] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.
[5] M. Roth, J. Stapel, R. Happee, and D. M. Gavrila, "Driver and pedestrian mutual awareness for path prediction and collision risk estimation," *IEEE Trans. Intell. Veh.*, vol. 7, no. 4, pp. 896–907, Dec. 2022.
[6] S. Gupta, M. Vasardani, and S. Winter, "Negotiation between vehicles and pedestrians for the right of way at intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 888–899, Mar. 2019.
[7] S. Martin, K. Yuen, and M. M. Trivedi, "Vision for intelligent vehicles & applications (VIVA): Face detection and head pose challenge," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2016, pp. 1010–1014.
[8] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 1–13.
[9] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1188–1197.
[10] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5738–5746.
[11] J. Levinson et al., "An analysis of SVD for deep rotation estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22554–22565.
[12] M. Roth and D. M. Gavrila, "DD-Pose - A large-scale driver head pose benchmark," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2019, pp. 927–934.
[13] M. D. Shuster, "A survey of attitude representations," *J. Astronautical Sci.*, vol. 41, no. 4, pp. 439–517, 1993.
[14] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7617–7627.
[15] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, pp. 1035–1046, 2019.
[16] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6D rotation representation for unconstrained head pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2022, pp. 2496–2500.
[17] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[18] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," *Pattern Recognit.*, vol. 127, 2022, Art. no. 108591.
[19] D. Derkach, A. Ruiz, and F. M. Sukno, "Tensor decomposition and nonlinear manifold modeling for 3D head pose estimation," *Int. J. Comput. Vis.*, vol. 127, no. 10, pp. 1565–1585, 2019.
[20] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[21] L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 157–171, Jan. 2021.
[22] F.-J. Chang et al., "Deep: Landmark-free FAME: Face alignment, modeling, and expression estimation," *Int. J. Comput. Vis.*, vol. 127, no. 6-7, pp. 930–956, 2019.
[23] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2065–2081, Sep. 2020.
[24] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognit.*, vol. 94, pp. 196–206, 2019.
[25] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2187–2196.
[26] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1087–1096.
[27] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image Vis. Comput.*, vol. 93, 2020, Art. no. 103827.
[28] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead – A large-Scale driver head pose dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1165–1174.
[29] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.
[30] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 596–609, Mar. 2020.
[31] J. M. D. Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker, "Fusion of keypoint tracking and facial landmark detection for real-time head pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 2028–2037.
[32] L. Sheng, J. Cai, T.-J. Cham, V. Pavlovic, and K. N. Ngan, "Visibility constrained generative model for depth-based 3D facial pose tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1994–2007, Aug. 2019.
[33] J. Gu, X. Yang, S. D. Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1531–1540.
[34] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López, "A reduced feature set for driver head pose estimation," *Appl. Soft Comput.*, vol. 45, pp. 98–107, 2016.
[35] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 17–24.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[37] A. Schwarz, Z. Lin, and R. Stiefelhagen, "HeHOP: Highly efficient head orientation and position estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–8.
[38] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "Deep head pose estimation from depth data for in-car automotive applications," in *Proc. Understanding Hum. Activities Through 3D Sensors*, 2018, pp. 74–85.
[39] B. Ahn, D. G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robot. Auton. Syst.*, vol. 103, pp. 1–12, 2018.
[40] A. Firintepe, M. Selim, A. Pagani, and D. Stricker, "The more, the merrier? A study on In-Car IR-based head pose estimation," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 1060–1065.
[41] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[43] V. Peretroukhin, M. Giamou, W. N. Greene, D. Rosen, J. Kelly, and N. Roy, "A smooth representation of belief over SO(3) for deep rotation learning with uncertainty," in *Proc. Robot.: Sci. Syst.*, Corvalis, Oregon, USA, 2020, pp. 1–9.

Markus Roth received the Diploma in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2014. Since then he has been working toward the Ph.D. degree with the Delft University of Technology, Delft, The Netherlands. He is also currently with Mercedes-Benz Research and Development, Perception and Maps Department, Stuttgart, Germany. His research interests include machine learning and video analysis for driver analysis, with a focus on driver head pose estimation and joint awareness between driver and pedestrians.

Dariu M. Gavrila (Member, IEEE) received the Ph.D. degree in computer science from the University of Maryland at College Park, College Park, MD, USA, in 1996. From 1997 to 2016, he was with Daimler R&D, Ulm, Germany, where he became a Distinguished Scientist. In 2016, he moved to Delft University of Technology, where he since heads the Intelligent Vehicles Group as a Full Professor. His research interests include sensor-based detection of humans and analysis of behavior, most recently in the context of the self-driving car in complex urban traffic. He was the recipient of the Outstanding Application Award 2014 and the Outstanding Researcher Award 2019, from the IEEE Intelligent Transportation Systems Society.