

Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation

Thorsten Hempel¹, Ahmed A. Abdelrahman², and Ayoub Al-Hamadi¹

Abstract—Estimating the head pose of a person is a crucial problem for numerous applications that is yet mainly addressed as a subtask of frontal pose prediction. We present a novel method for unconstrained end-to-end head pose estimation to tackle the challenging task of full range of orientation head pose prediction. We address the issue of ambiguous rotation labels by introducing the rotation matrix formalism for our ground truth data and propose a continuous 6D rotation matrix representation for efficient and robust direct regression. This allows to efficiently learn full rotation appearance and to overcome the limitations of the current state-of-the-art. Together with new accumulated training data that provides full head pose rotation data and a geodesic loss approach for stable learning, we design an advanced model that is able to predict an extended range of head orientations. An extensive evaluation on public datasets demonstrates that our method significantly outperforms other state-of-the-art methods in an efficient and robust manner, while its advanced prediction range allows the expansion of the application area. We open-source our training and testing code along with our trained models: <https://github.com/thohemp/6DRepNet360>.

Index Terms—Head pose estimation, full range of rotation, rotation matrix, 6D representation, geodesic loss.

I. INTRODUCTION

HEAD pose estimation follows the objective of predicting the human head orientation from images and is a crucial step in many computer vision algorithms. Applications are wide-ranging and include attention estimation [1], [2], [3], face recognition [4], [5], and the estimation of facial attributes [6], [7], which again are vital features in driver assistance systems [8], [9], [10], augmented reality [11], [12], and human-robot interaction [13], [14], [15]. The vast majority of present methods [16], [17], [18], [19], [20], [21], [22], [23] narrow down the research issue to the estimation of solely frontal poses with a limited rotation range. This favors the leverage of the facial feature-richness and suitable, widely available training datasets. However, in uncontrolled application scenarios [24], [25], [26] head orientations are likely to surpass the narrow angle range that most methods are trained for and, consequently, produce random and

inaccurate head pose predictions. In view of extending the prediction to the full area of rotation range, the current state of research is challenged by two key limitations. The first is the absence of comprehensive datasets that cover the full range of head orientations [27]. The second equally decisive and often neglected factor is an appropriate rotation representation, as it significantly impacts the model's ability to effectively learn the connection between visual pose appearance and corresponding parameterization [28]. For instance, the commonly used Euler angle and quaternion representation suffer from ambiguity and discontinuity problems that lead to an unstable training process and a mediocre prediction performance if plainly applied [16], [19], [23], [29]. This behavior even intensifies for stronger rotations in the narrow range spectrum.

We overcome these limitations by proposing a rotation matrix-based 6D representation for efficient and unconstrained network training that we further enhance with a geodesic based loss. Additionally, we take up the ambitious challenge of predicting the full range of rotation by agglomerating new training data with enhanced pose variation. For this matter, we utilize the CMU Panoptic [30] dataset and apply an automatic head pose labeling process to generate head pose samples with focus on the back of the head. We combine these samples with the popular 300W-LP [31] head pose dataset and, together, receive a large scaled dataset with greatly expanded head rotation variations. Finally, the training of our proposed model on this new agglomerated data enables us to predict a significantly extended range of head orientations. We examine our approach in multiple experiments on public datasets that testify our method state-of-the-art accuracy and remarkable robustness in predicting challenging poses. At the same time, it is able to handle a many times greater range of head pose orientations compared to current methods from the literature. Fig. 1 shows examples of orientation predictions from this model for versatile head poses. To the best of our knowledge, we are the first to tackle the full range of head pose estimation in this extensive and conclusive way. In summary, we make the following contributions:

- We introduce a simplified and efficient 6-parameter rotation matrix representation for regressing accurate head orientations without suffering ambiguity problems.
- We propose a geodesic distance approach for network penalizing to encapsulate the training loss within the Special Orthogonal Group $SO(3)$ manifold geometry.
- We utilize the CMU Panoptic dataset [30] to expand the traditional 300W-LP [31] head pose dataset with full rotation head pose appearance.

Manuscript received 18 August 2022; revised 2 February 2024; accepted 29 February 2024. Date of publication 21 March 2024; date of current version 26 March 2024. This work was supported in part by the Federal Ministry of Education and Research of Germany (BMBF) Project AutoKoWaT under Grant 13N16336 and in part by German Research Foundation (DFG) Project under Grant AL 638/15-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mohammed Bennamoun. (Corresponding author: Thorsten Hempel.)

The authors are with the Faculty of Electrical Engineering and Information Technology, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany (e-mail: Thorsten.Hempel@ovgu.de).

Digital Object Identifier 10.1109/TIP.2024.3378180

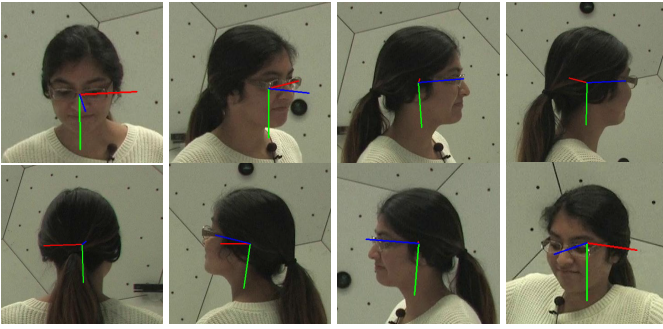


Fig. 1. Example images of predicted orientations of various rotated heads.

- We create a new head pose prediction model that surpasses the prediction range of current methods and at the same time achieves lower errors on common test datasets.
- We demonstrate the superiority of our approach in accuracy and robustness in multiple experimental setups.
- We conduct an ablation study to evaluate the impact of each component of our model on the achieved results.

Fig. 2 shows an overview of our proposed method. Each component will be explained in detail in the following sections. Inspired by the 6D representation that is used in our approach, we call our network 6DRepNet. An earlier version of this work was published in [32], where we presented an initial approach for 6D-based narrow angle prediction. In this version, we enhance this previous work with an improved training procedure, propose an approach for tackling the prediction of the full range of orientation, and provide a more detailed model including an extensive comparison with the state-of-the-art, error analysis and ablation studies.

Our training, testing code, and trained models are made publicly available to facilitate research experimentation and practical application development.

II. RELATED WORK

In recent years, facial analysis along with vision-based head orientation prediction emerged with the rise of neural networks. Current methods are commonly divided into landmark-based and landmark-free approaches. Landmark-based methods [33], [34], [35], [36] detect facial landmarks as a primary step and subsequently recover the 3D head pose by aligning the predicted landmarks with a standardized 3D head model [37], [38]. Under ideal circumstances, this approach can lead to very accurate head orientation estimations, but it is highly dependent on the precise predictions of the landmark positions. Also, it requires the target head to be shaped similar to the head model to achieve an accurate alignment. Other methods surmount these constraints by directly predicting 3D facial landmarks, from which the head pose can be straightforwardly determined based on the localized landmarks [39], [40], [41]. However, 3D landmarks and their 2D counterparts are only located in the facial area. Head poses with significant occlusions and particularly strong rotations only reveal little or no visible facial area, making landmark-based methods more prone to failure [42], [43]. Landmark-free approaches overcome these limitations by directly estimating the head pose from the images in an end-to-end fashion. These methods

commonly use deep neural networks to formulate the orientation prediction as an appearance-based task. As one of the first of its kind, HopeNet [16] presented an RvC [44] approach by binning the target angle range to combine a cross-entropy and a mean squared error loss function for Euler angle prediction. Along with this classification approach, they at the same time reduced the predictable rotation range within ± 99 degrees for yaw, pitch, and roll. Later, QuatNet [19] adapted the cross-entropy paradigm with limited prediction range and proposed to split classification and regression into separate network branches. One branch is used for classifying the Euler angles and the second one regresses the pose in quaternion representation. Similarly, HPE [18] treats classification and regression separately and averages the outputs as a pose regression subtask. WHENet [29] keeps the single branch strategy, switches to an EfficientNet [45] backbone and increases the number of bins for the yaw network branch to extend the predictable angle range. Whereas FSA-Net [17] proposes a network with a stage-wise regression and feature aggregation scheme for predicting Euler angles. TriNet [46] adapts this method, but estimates the three unit vectors of the rotation matrix instead of Euler angles and incorporates an additional orthogonality loss to stabilize the predictions. MFD-Net [21] likewise follows the rotation matrix representation but uses its Fisher distribution to model rotation uncertainty and to find its maximum likelihood. Another probabilistic approach was proposed by Liu et al. [47] who train on Gaussian label distributions. Whereas FDN [20] targets optimized feature extraction by proposing a feature decoupling method to explicitly learn discriminative features of different head orientations. DDD-Pose [22] seeks to diversify the training data by proposing an advanced augmentation scheme. The current state-of-the-art results are achieved by RankPose [23] closely followed by MNN [48]. RankPose uses paired training samples to introduce a ranking loss for penalizing incorrect ordering of the Euler pose estimation. MNN and img2pose [49] predict the rigid transformation between the head and the camera.

In general, frequent approaches in the area of head pose estimation achieved continuous improvement over the recent years, yet they still lack of comprehensive solutions for predicting the full range of head pose rotation. First, it became a common convention to split up the continuous rotation variables into bins to convert the problem into a classification task in order to stabilize the predictions [16], [18], [19], [20], [29]. However, this is problematic as pruning segments of angles into bins will consequently lead to a loss of information. Apart from that, this constraining approach is commonly combined with reducing the target space [16], [18], [19], [20], [29] which eliminates the opportunity of tackling full rotation estimation. A few works overcome these limiting factors by using the rotation matrix as a more suitable rotation representation [21], [46], [50], but neither deal with more efficient ways of regression nor address its potential for expanding the prediction range. As a consequence, the area of full head pose prediction is still rarely explored yet. WHENet [29] was one of the first to approach full yaw prediction by extending the bin range for the yaw angle and proposing a wrapping loss to handle the influence of the gimbal lock. However, their

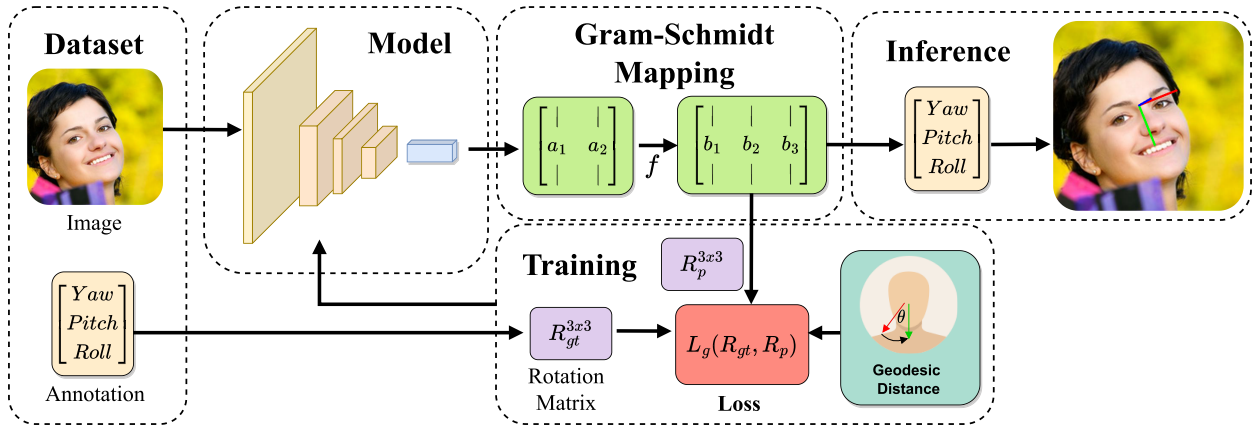


Fig. 2. Overview of the proposed method.


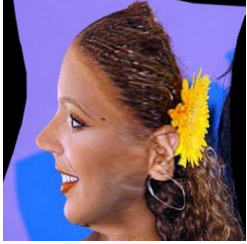
	
Euler Angles	Euler Angles
$[87.73 \quad 89.32 \quad -87.93]$	$[-92.51 \quad -98.02 \quad 81.73]$
Quaternions	Quaternions
$[0.707 \quad -0.023 \quad -0.707 \quad 0.031]$	$[-0.707 \quad -0.029 \quad 0.706 \quad 0.041]$
Rotation matrix	Rotation matrix
$\begin{bmatrix} 0.000 & 0.012 & -0.999 \\ -0.076 & -0.997 & -0.012 \\ -0.997 & 0.076 & 0.000 \end{bmatrix}$	$\begin{bmatrix} 0.002 & -0.017 & -0.999 \\ 0.100 & -0.995 & 0.017 \\ -0.995 & -0.100 & -0.001 \end{bmatrix}$

Fig. 3. Data samples from 300W-LP dataset with different rotation parameterization.

method still tightly restricts pitch and roll within ± 99 degrees. The same restriction is applied by Viet et al. [50] in their multitask approach, where they face detection and head pose estimation. As rotation representation, they use the rotation matrix and follow the same computational extensive approach as TriNet [46] to obtain orthogonality.

III. METHOD

In the following, we will give details about our proposed method. We start with preliminary information about different rotation representations. Based on its insights, we propose a rotation parametrization scheme to overcome the limitation of the related works. As an accompanying measure, we will introduce a geodesic distanced based loss to precise and stabilize the network penalty for training.

A. Preliminaries

In general, the orientation of a rigid body in the three-dimensional space can be described by multiple kinds of mathematical representation. The most common and widely used one is the Euler angle representation that is used to describe the rotation around each axis of the coordinate system

(typical denoted as *yaw*, *pitch*, *roll*). Despite its intuitiveness, Euler angles face limitations when it comes to the specific orientation state, where the second elemental rotation reaches 90 or -90 degrees. Given this setup, *yaw* and *pitch* align on the same plane and create infinitive solutions for the same rotation state. This behavior is known as *gimbal lock* as the first and third axis are locked under this particular condition. The gimbal lock represents the extreme case for the limitations of Euler angles. However, the dependency between first and third angle is a fundamental property of Euler angles, that just becomes stronger the more the pitch reaches the gimbal lock state. As a consequence, the Euler angle representation does not behave in the same continuous form as its visual appearance counterpart that has a detrimental impact on the performance of neural networks.

Another type of orientation is called axis-angle representation, which consists of a unit vector $v = (\tilde{x}, \tilde{y}, \tilde{z})$ that defines the axis of the rotation and an angle θ that describes the magnitude of its rotation. Closely related to the axis-angle representation, another type called rotation quaternions q with also four parameters q_0, q_1, q_2, q_3 can be derived by $q_0 = \cos(\frac{\theta}{2})$, $q_1 = \tilde{x} \sin(\frac{\theta}{2})$, $q_2 = \tilde{y} \sin(\frac{\theta}{2})$, $q_3 = \tilde{z} \sin(\frac{\theta}{2})$. Quaternions and the axis-angle representation are not affected by the gimbal lock, but they still have an ambiguity that is introduced by their antipodal symmetry with $-v = v$ and $-q = q$, respectively. As a result, every orientation can be described by two different representations that are maximum far apart. A more comprehensive notation is the rotation matrix $R^{3 \times 3}$ that consists of 9 parameters. Despite its increased number of parameters, it comes with the crucial advantage that it provides a continuous representation with a unique parameterization for each rotation. Fig.3 shows an example of two dataset samples with similar pose appearances. Yet, their Euler angle and quaternion ground truth are parameterized very differently. Only the rotation matrices reflects the similarity in the pose appearance. In $SO(3)$ the matrix representation R is sized 3×3 with an orthogonality constraint $RR^T = I$, where R^T is the transposed matrix and I the identity matrix.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \quad (1)$$

One could now try to regress the rotation matrix directly, but this would require finding all nine parameters that at the same time satisfy the orthogonality constraint. The orthogonality can also be enforced in a sequential step by either using the Gram-Schmidt process or the singular value decomposition (SVD). The SVD is an extensive approach for finding those orthogonal vectors that are the nearest to the predictions. The Gram-Schmidt method requires discarding one vector in order to recreate the orthogonal matrix from the remaining two.

B. 6D Representation

In section III-A we show that a key aspect for tackling direct orientation predictions is the use of an appropriate rotation representation that is unambiguously interpretable by neural networks. For this matter, we use the rotation matrix representation as a superior alternative to Euler angles, quaternions, and axis-angles. Inspired by Zhou et al. [28], we satisfy the orthogonality constraint by performing the Gram-Schmidt mapping inside the representation itself, which avoids extensive post-processing. We simply drop the last column vector of the rotation matrix that reduces the 3×3 matrix into a 6D rotation representation

$$g_{GS} = \left(\begin{bmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{bmatrix} \right) = \begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix}, \quad (2)$$

which has been reported to introduce smaller errors for direct regression [28]. Then, the predicted 6D representation matrix is mapped back into $SO(3)$ with

$$f_{GS} = \left(\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \right) = \begin{bmatrix} | & | & | \\ b_1 & b_2 & b_3 \\ | & | & | \end{bmatrix}, \quad (3)$$

where the resulting column vectors are defined as

$$\begin{aligned} b_1 &= \frac{a_1}{\|a_1\|}, \\ b_2 &= \frac{u_2}{\|u_2\|} \text{ with } u_2 = a_2 - (b_1 \cdot a_2)b_1, \\ b_3 &= b_1 \times b_2. \end{aligned} \quad (4)$$

Hereby, the last column vector is simply determined by the cross product that ensures that the orthogonality constraint is satisfied for the resulting 3×3 matrix:

As a result, our network has only to predict 6 parameters that are mapped into a 3×3 rotation matrix in a subsequent transformation process that incorporates the orthogonality constraint as well.

C. Geodesic Loss

The l_2 -norm is the commonly used loss function for head pose related tasks. However, using the Frobenius norm for measuring distances between two matrices would break with the $SO(3)$ manifold geometry. Instead, the shortest path between two 3D rotations is geometrically interpreted as the geodesic distance. Let R_p and $R_{gt} \in SO(3)$ be the estimated and the ground truth rotation matrices, respectively, then the

geodesic distance between both rotation matrices is defined as:

$$L_g = \cos^{-1} \left(\frac{\text{tr}(R_p R_{gt}^T) - 1}{2} \right). \quad (5)$$

In the following, we will use this metric as a loss function for our neural network to compute accurate distance information between the predicted and ground truth orientation.

IV. EXPERIMENTS

We perform an extensive evaluation of our method. We begin the specification of our used datasets, evaluation metrics and implementation setup, followed by a comprehensive comparison with other state-of-the-art methods in cross-dataset and intra-dataset tests. Further analysis includes a detailed error analysis and ablation studies on used loss functions and backbones.

A. Datasets

We conduct our evaluation with the aid of different kinds of data. The most common and public available datasets are 300W-LP [31], AFLW2000 [51], and BIWI [52].

300W-LP: 300W-LP consists of 66,225 face samples collected from multiple databases including LFPW [53], AFW [54], HELEN [55] and iBUG [56] that are further enhanced to 122,450 samples by image flipping. It is based on around 4000 real images. The ground truth is provided in the Euler angle format. For training, we convert them into the matrix form.

AFLW2000: The AFLW2000 dataset contains the first 2,000 images from the ALFW dataset annotated with the ground truth 3D faces and the corresponding 68 landmarks. It contains samples with large variations, different illumination, and occlusion conditions.

BIWI: The BIWI dataset includes 15,678 images that were created in a lab environment with 20 participants. In this dataset, the head takes up only a small area in the images. Hence, we use the MTCNN [57] face detector to loosely crop the heads from the images. All of the above listed datasets provide, due to their nature of annotation, only samples with a frontal view of the faces (mostly between -99° and $+99^\circ$ range of yaw). Therefore, they cannot be used for the training of the entire head orientation range.

CMU Panoptic: Therefore, we utilize another dataset called CMU Panoptic [30] that makes it possible to generate annotated head images with full rotation appearances. In this dataset, a variety of subjects perform arbitrary tasks inside a dome, that is equipped with 31 evenly arranged HD cameras. The main focus of this dataset is to capture the subject's poses, but it also provides 3D facial landmark annotation and camera intrinsics and extrinsics. This enables to extract head pose annotation from all the different camera angles, that was initially harnessed by Zhou and Gregson [29]. There are 30 sequences public available with multiple subjects per scene, that are standing in a ring with each subject being oriented towards the center of the dome. When extracting the head crops with only accepted those with a minimum size of 320 for

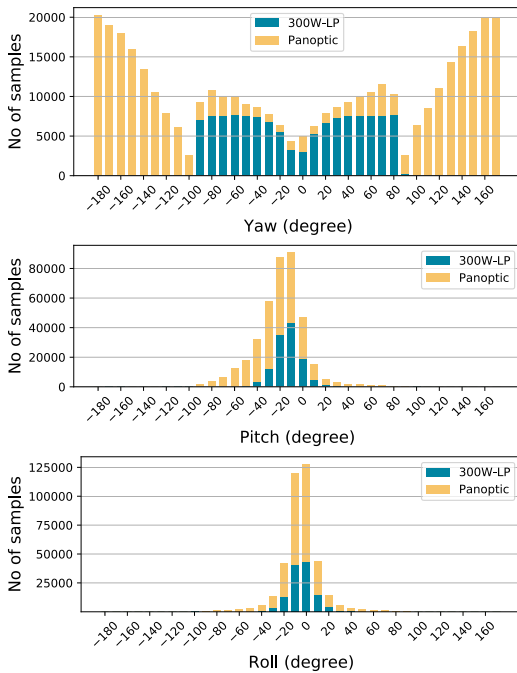


Fig. 4. Dataset distribution.

both axis, which gives us a dataset with 113914 samples in total. Because of the subject’s spacial setup, the majority of the samples are ones showing the back of the head. Samples with frontal face view are likelier to be sorted out by too small sized, as these face images were taken from more far distance. Therefore, we create a combination of the 300W-LP and the CMU Panoptic dataset that includes 236,364 data samples spanning the entire range of yaw rotation. The range of pitch is slightly expanded as we also use the samples that are generated from cameras attached to the ceiling of the CMU Panoptic dome. The distribution of this new training data is shown in Fig. 4. It should be noted that we use the Euler angles for presentation purposes that cannot exactly represent the distribution of visual appearance in the dataset, as discussed in section III-A.

B. Evaluation Metrics

We use two different evaluation metrics to quantify the head pose estimations error. The first one is the most common Mean Absolute Error (MAE) of the Euler angles,

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N (|x_g - x_p|), \quad (6)$$

where N is the number of face images and x_g and x_p represent the ground truth and predicted pose parameters, respectively. Secondly, we calculate the Mean Absolute Error of the vectors (MAEV) of the rotation matrix. This metric was introduced by [46] in order to surpass the limitations of the Euler representation and to provide a more meaningful picture of the appearance differences between predicted and ground truth orientation. The MAEV defines the angle error between the

three vectors of the rotation matrix,

$$\text{MAEV} = \frac{1}{N} \sum_{i=1}^N \cos^{-1} \left(\frac{v_g \cdot v_p}{|v_g| |v_p|} \right), \quad (7)$$

where N , again, is the number of face images in the dataset and v_g and v_p are the ground truth and the predicted head orientation vectors.

C. Implementation Details

We implement our proposed network using PyTorch [58]. As backbone, we choose ResNet50 [59] to enable a fair comparison with other methods [16], [19], [22], [23], [46], that chose the same feature extractor. The backbone’s weights are pretrained with the ImageNet [60] dataset. For the final layers, we choose a single fully connected layer with 6 outputs. The network is trained for 80 epochs with a batch size of 80 using the Adam optimizer with a learning of $1e^{-4}$. To exploit full generalization potential, we also extensively augment our training data using Albumentations [61] by applying random horizontal flipping, random scaling and cropping, random rotation up to $[-45, +45]$ degrees, random occlusions, and further image color operations including random blur, random brightness contrast changes, and random RGB shifts.

D. Comparison With State-of-the-Art

In this section, we conduct a comprehensive comparison with the state-of-the-art. We start with a cross-dataset evaluation to analyze our model’s generalization capabilities, followed by an intra-dataset experiment and a detailed error analysis for further performance assessment.

1) *Cross-Dataset Evaluation*: In our first experiment, we want to evaluate our approach against the state-of-the-art methods. To the end, we train two models. The first model (*6DRepNet*) will strictly follow the common training convention by using the synthetic 300W-LP dataset for training and the two real-world datasets AFLW2000 and BIWI for testing. This will provide comparable information about our method’s performance of directly regressing a diminished rotation matrix. For another model (*6DRepNet360*) we change the training setup by replacing the 300W-LP dataset for training with our combined dataset (CMU + 300W-LP) for full rotation appearance training. The remaining training configuration will remain the same to place the focus on the impact of the enriched training data for the evaluation. We provide numerical results in Mean Absolute Error (MAE) of the Euler angles and in Mean Absolute Error of the rotation matrix vectors (MAEV). Most of the current state-of-the-art methods don’t provide the MAEV for their results, so we retest these methods that have been open sourced and calculate the MAEV based on the converted rotation matrices for each test sample.

Table I shows the results from the two model setups along with the results from other methods from the recent literature. For better interpretation, we added an extra column (R) to show which methods are trained to predict a larger range of rotations and which ones restrict their predictions to frontal poses. From the 15 listed methods, only two approached the exceeding of narrow angle range head pose estimation.

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE AFLW2000 AND BIWI DATASET. ALL MODELS ARE TRAINED ON THE 300W-LP DATASET. RESULTS FROM METHODS WITH POSITIVE I ARE GENERATED BY OUR OWN TESTS. METHODS WITH NEGATIVE I ARE NOT OR ONLY PARTIALLY OPEN-SOURCE. THEIR RESULTS ARE CLAIMS FROM AUTHORS. METHODS WITH POSITIVE R TARGET THE PREDICTION OF A WIDER RANGE OF ROTATION

	Euler										Vector							
	AFLW2000					BIWI					AFLW2000				BIWI			
	I	R	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE	Left	Down	Front	MAEV	Left	Down	Front	MAEV
HopeNet [16]	✓	✗	6.40	6.53	5.39	6.11	4.54	5.15	3.37	4.36	7.98	6.40	8.54	7.64	6.21	5.75	7.05	6.33
FSA-Net [17]	✓	✗	4.83	6.25	4.94	5.34	4.64	5.61	3.57	4.61	6.88	6.52	7.28	6.89	6.27	6.29	7.38	6.65
HPE [18]	✗	✗	4.80	6.18	4.87	5.28	3.12	5.18	4.57	4.29	-	-	-	-	-	-	-	-
QuatNet [19]	✗	✗	3.97	5.62	3.92	4.50	2.94	5.49	4.01	4.15	-	-	-	-	-	-	-	-
TriNet [46]	✓	✗	4.36	5.81	4.51	4.89	3.11	5.09	5.20	4.47	6.16	5.95	6.82	6.31	6.58	5.80	7.55	6.64
WHENet-V [29]	✗	✗	4.44	5.75	4.31	4.83	3.60	4.10	2.73	3.48	-	-	-	-	-	-	-	-
WHENet [29]	✗	✓	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81	-	-	-	-	-	-	-	-
FDN [20]	✗	✗	3.78	5.61	3.88	4.42	4.52	4.70	2.56	3.93	-	-	-	-	-	-	-	-
Viet et al [50]	✗	✓	-	-	-	-	4.62	4.29	4.52	4.48	-	-	-	-	-	-	-	-
MFDNet [21]	✗	✗	4.30	5.16	3.69	4.38	3.40	4.68	2.77	3.62	-	-	-	-	-	-	-	-
DDD-Pose [22]	✗	✗	4.38	4.85	3.44	4.22	4.60	6.02	2.94	4.52	-	-	-	-	-	2.94	-	-
Liu et al. [47]	✗	✗	3.03	5.06	3.68	3.93	4.12	5.61	3.15	4.29	-	-	-	-	-	-	-	-
img2pose [49]	✓	✗	3.42	5.03	3.28	3.91	4.56	3.54	3.25	3.78	6.00	5.20	6.55	5.92	4.83	5.28	6.04	5.38
MNN [48]	✗	✗	3.34	4.69	3.48	3.83	3.98	4.61	2.39	3.66	-	-	-	-	-	-	-	-
RankPose [23]	✓	✗	3.26	4.72	3.23	3.74	4.54	5.61	3.05	4.40	4.40	4.42	5.08	4.63	5.81	5.91	7.39	6.37
6DRepNet	✗		3.27	4.58	2.98	<u>3.61</u>	3.23	5.32	2.78	3.78	4.33	4.17	5.06	<u>4.52</u>	4.66	5.29	6.03	5.32
6DRepNet360	✓		3.73	5.52	3.53	4.26	3.37	3.87	2.93	<u>3.39</u>	5.18	4.70	6.04	<u>5.31</u>	4.64	4.57	5.34	4.85

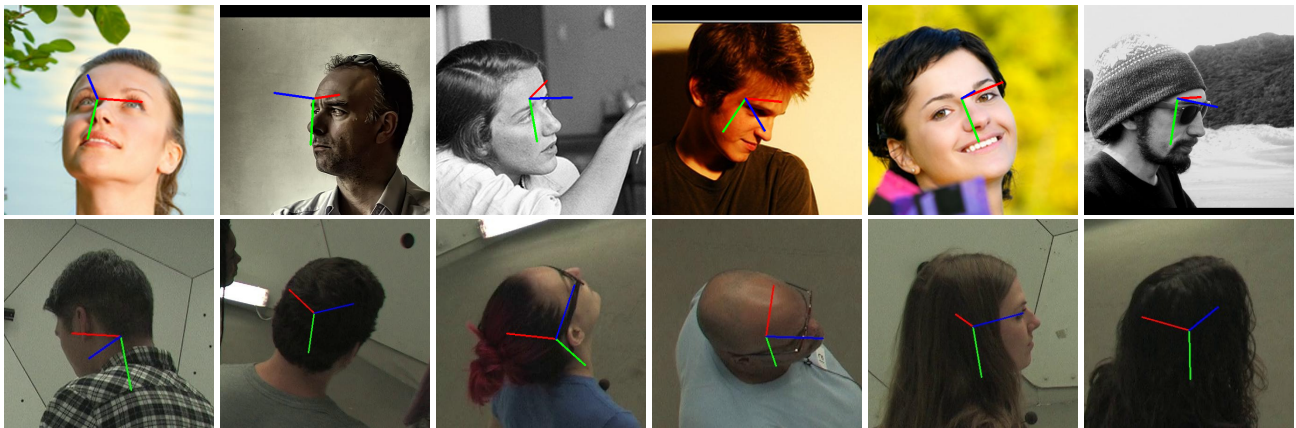


Fig. 5. Example images with converted Euler angle visualization from the AFLW2000 dataset (first row) and the CMU Panoptic test dataset (second row).

a) *6DRepNet*: The table demonstrates that our model that was solely trained on the 300W-LP dataset outperforms all other methods on the AFLW2000 test dataset and surpasses the current top performer RankPose on AFLW2000 in Euler and vector errors. Besides the overall error rate, our model achieves top performing results for the pitch and roll error and equal results to the best reported yaw error. This indicates a very stable network learning, resulting in robust prediction properties. On the BIWI dataset, it achieves competitive results in respect to MAE and best results in respect to MAEV. The latter ought to be considered with caution, as there are no MAEV results reported for the MAE top performers.

b) *6DRepNet360*: Our second model, 6DRepNet360, achieves very competitive results on AFLW2000 and even new state-of-the-art results on BIWI by surpassing WHENet-V by 3%. Noticeably, this model only differs in its training data, where the added data aims to expand the predictable detection range of the yaw rotation. Yet, these samples include numerous stronger pitch rotations than 300W-LP (see Fig. 4). We argue that these samples benefit the model's performance

for processing the challenging poses from the BIWI dataset, as the error for the pitch is reduced by 33% compared to our the solely on 300W-LP trained model. Remarkably, WHENet is also trained for wide yaw predictions and is therefore most suitable to compare it with our 6DRepNet360. While WHENet is reported to perform even worse than its 300W-LP equivalent WHENet-V, our 6DRepNet360 model achieves over 20% lower error rates on AFLW2000 and over a 10% higher accuracy on BIWI. We believe that our choice of the 6D rotation matrix as rotation representation instead of WHENets Euler angle has a major impact on our superior results. In terms of rotation representation, TriNet is the most similar method to ours. But in contrast to our 6 parameter approach, they predict the entire 9 parameter rotation matrix and use an SVD to find an orthogonal-constrained solution. We argue that our more efficient approach leads to a higher reported accuracy.

Fig. 5 shows qualitative results from our 6DRepNet360 model. The first row illustrates prediction on test images from the AFLW2000 dataset with strong varieties of background,

TABLE II

EULER ERROR COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE70/30 BIWI DATASET. RESULTS FROM METHODS WITH POSITIVE I ARE GENERATED BY OUR OWN TESTS. METHODS WITH NEGATIVE I ARE NOT OR ONLY PARTIALLY OPEN-SOURCE. THEIR RESULTS ARE CLAIMS FROM AUTHORS

BIWI Euler					
	I	Yaw	Pitch	Roll	MAE
HopeNet [16]	✓	3.35	4.66	3.00	3.67
FSA-Net [17]	✓	6.79	9.18	4.56	6.84
FDN [20]	✗	3.00	3.98	2.88	3.29
MDFNet [21]	✗	2.99	3.68	2.99	3.22
TriNet [46]	✓	2.79	3.28	2.53	2.87
DDD-Pose [22]	✗	3.04	2.94	2.43	2.80
6DRepNet		2.39	2.96	2.05	2.47

TABLE III

VECTOR ERROR COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE70/30 BIWI DATASET. RESULTS FROM METHODS WITH POSITIVE I ARE GENERATED BY OUR OWN TESTS. METHODS WITH NEGATIVE I ARE NOT OR ONLY PARTIALLY OPEN-SOURCE. THEIR RESULTS ARE CLAIMS FROM AUTHORS

BIWI Vector					
	I	Left	Down	Front	MAEV
FSA-Net [17]	✓	9.09	10.19	11.26	10.18
HopeNet [16]	✓	5.55	5.64	5.78	5.66
TriNet [46]	✓	4.12	4.47	4.24	4.28
6DRepNet		3.39	3.27	3.89	3.52

TABLE IV

MODEL PERFORMANCE ON THE CMU PANOPTIC + 300W-LP COMBINED DATASET. 70% OF THE DATASET IS USED FOR TRAINING AND THE REMAINING 30% FOR TESTING. RESULTS FROM METHODS WITH POSITIVE I ARE GENERATED BY OUR OWN TESTS. METHODS WITH NEGATIVE I ARE NOT OR ONLY PARTIALLY OPEN-SOURCE. THEIR RESULTS ARE CLAIMS FROM AUTHORS

CMU Panoptic					
	I	Yaw	Pitch	Roll	MAE
Viet et al. [50]	✗	9.55	11.29	8.32	9.72
WHENet [29]	✗	8.51	7.67	6.78	7.66
6DRepNet360		2.08	3.16	2.75	2.66

lightning, and camera angle. The second row shows test results with very strong head rotations from the CMU Panoptic test set that exceed the common pm 99 degrees restrictions. In contrast to AFLW2000, it is captured in a laboratory environment with consistent lightning conditions and background. Nevertheless, 6DRepNet robustly predicts the head poses from varying camera angles. A very noteworthy example is the rightmost test image, as it presents a very challenging instance. While for frontal faces even stronger rotated poses provide meaningful features, visual cues are in this example mainly restricted to the head's shape. Yet, our model is able to predict reliable orientations even for these challenging kind of head poses.

2) Intra-Dataset Evaluation:

a) *BIWI*: In a second experiment, we follow the convention by FSA-Net [17] and randomly split the BIWI dataset in

TABLE V

ANALYSIS OF THE INFLUENCE OF DIFFERENT LOSS FUNCTIONS L_{MSE} AND GEODESIC LOSS L_g ON THE MAE

AFLW2000								
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
L_{MSE}	3.38	4.89	3.33	3.87	3.19	6.52	2.81	4.17
$L_g + L_{MSE}$	3.26	4.65	3.09	3.67	3.17	5.69	2.76	3.88
L_g	3.27	4.58	2.98	3.61	3.23	5.32	2.78	3.78

a ratio of 7:3 for training and testing, respectively. Table II and Table V show our results compared with other state-of-the-art methods that followed the same testing strategy. We retested those models, that provide source code information, for an additionally MAEV error report. The remaining results are claims by the authors. It demonstrates that our method outperforms all other methods by a margin of more to 10%. In terms of the individual rotation angles, our approach produces very consistent results by achieving the best results on yaw and roll, and equal results to the state-of-the-art DDD-Pose for the pitch angle. This supports the observed robustness in the cross-dataset evaluation and demonstrates, that achieving stable accurate results for all three angles does not only depend on the trained dataset, but rather on our proposed method itself. This is also reflected in Table V, where our approach achieves the best overall MAEV results as well as for each single vector.

b) *CMU panoptic + 300W-LP*: In a final experiment, we evaluate our model in an intra-dataset test on our combined dataset that comprises the data from the CMU Panoptic and the 300W-LP dataset. To this end, we randomly split the dataset into 70% training data and 30% test data. To the best of our knowledge, Viet et al. [50] and WHENet are the only methods that published test results on CMU Panoptic. However, [50]'s prediction pipeline additionally includes face detection and their test set comprises solely samples from CMU Panoptic. Therefore, the comparison ought to be considered with caution. More similar to our experimental approach, WHENet tests on a combination of CMU Panoptic and 300W-LP, but its size and composition are not specified. Thus, our results are mainly for future reference, and we will publish our test list to provide other methods with the capability of precise comparison.

3) *Error Analysis*: To receive a more detailed impression of our model performance, we conduct an error analysis with four other state-of-the-art methods (HopeNet, FSA-Net, RankPose, TriNet) where we split up the errors on the AFLW2000 of range $[-99^\circ, 99^\circ]$ into intervals of 33° . All models were solely trained on 300W-LP. The results are shown in Fig. 6 where each Euler angle is illustrated in a separate graph. It gives insight that in general the prediction error for all methods increases with stronger rotations. It is conspicuous, though, that this error increase is much lower for 6DRepNet compared to all the other methods, especially for the pitch and roll. While Table I shows that our model overall outperforms RankPose by 3%, this detailed error analysis illustrates that our 6DRepNet achieves over 60% smaller error rates for extreme pitch and roll rotations. This is yet another confirmation that our model does not only achieve state-of-that results, but at the same time

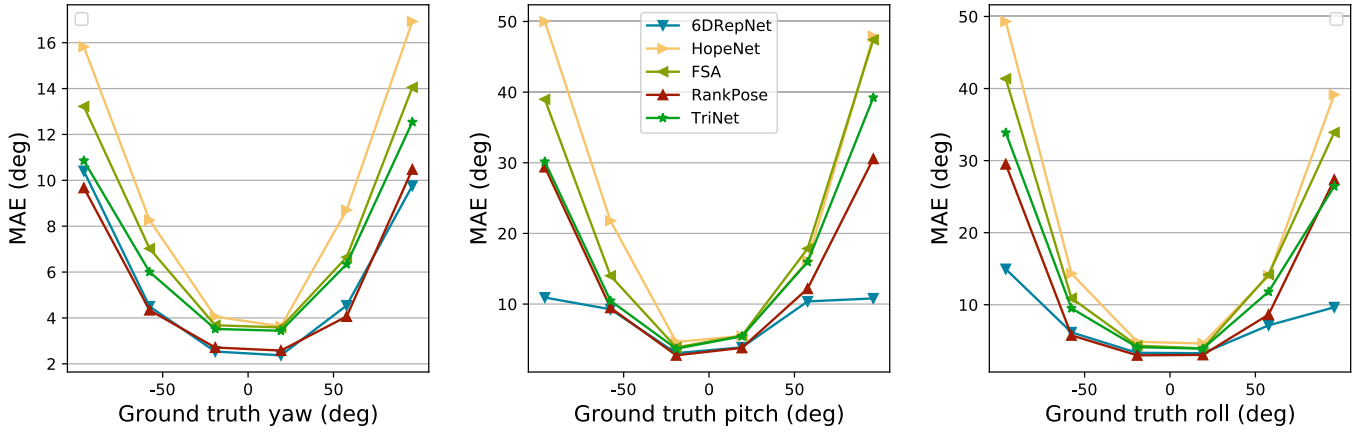


Fig. 6. Error analysis for angle intervals on the AFLW2000 dataset.

provides very robust predictions even in extremely challenging test cases. We argue that the significant error reduction for pitch and roll can be attributed to our model’s efficient learning capabilities, enabling accurate predictions even with a limited amount of large-angle training samples (see Fig. 4).

E. Ablation Study

In the following, we will analyze how each of our model’s remaining components impacts our reported results. This includes the backbone, that is responsible for the feature extraction, and our proposed loss function, which differs from other methods in the literature.

1) *Loss Function*: Most current methods use the Mean Squared Error (MSE)

$$L_{MSE} = \frac{1}{N} \sum_{i=0}^N (y_p - y_{gt})^2 \quad (8)$$

for calculating the loss in the training procedure. We argue that the geodesic distance gives a better feedback about the distance between prediction and ground truth and, thus, is better suited to be used as a loss function. To prove this, we conduct another experiment where we repeat our previous tests, but this time we train our network with the MSE distance loss and with a combination of MSE and the geodesic loss L_g (see Eq. 5). Table III shows these results compared to our models trained with geodesic distance loss. It states that the network with geodesic loss penalty performed significantly better than the one that used MSE and slightly better than the combination of MSE and L_g .

2) *Rotation Formalisms*: We evaluate the performance of various rotation formalism by using a ResNet50 backbone with single final fully connected layer. This layer comprises three output neurons for the Euler-based model, four neurons for the quaternion-based model, six for the 6D formalism and nine for the rotation matrix based model. All models are trained using MSE loss, except for an additional 6D-based model, which utilizes the proposed Geodesic distance-based loss. The results are presented in Table VI and indicate the highest error for the Euler angle-based model, followed by the quaternion- and rotation matrix-based models. The 6D

TABLE VI
ANALYSIS OF VARIOUS ROTATION FORMALISM

AFLW2000					
	Loss	Yaw	Pitch	Roll	MAE
Euler	L_{MSE}	9.57	6.35	5.16	7.03
Quaternion	L_{MSE}	6.33	6.18	5.07	5.86
Rotation Matrix	L_{MSE}	4.00	5.34	3.96	4.43
6D (proposed)	L_{MSE}	3.38	4.89	3.33	3.87
6D (proposed)	L_g (proposed)	3.27	4.58	2.98	3.61

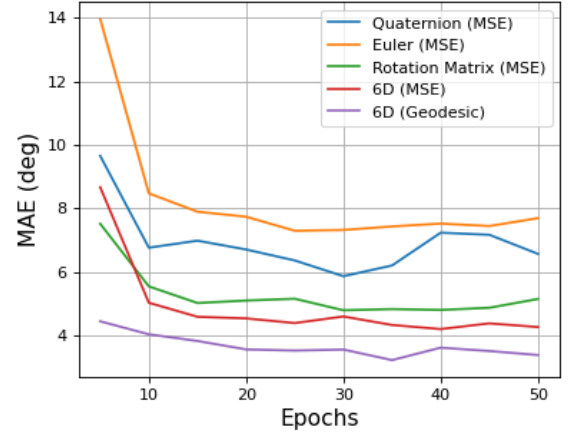


Fig. 7. Comparison of MAE for various rotation formalism based on a ResNet50 backbone and MSE loss. For the proposed 6D representation, an additional series with Geodesic loss is provided. All results are based on the AFLW2000 test set.

formalism-based models archive the best results, in which the Geodesic loss model outperforms the MSE model in Yaw, Pitch and Roll error rates. These findings support our claim that the 6D rotation representation in combination with the Geodesic-based loss facilitates efficient training and yields highly accurate rotation prediction models. To further analyze the training process, we illustrate the test performance of the trained models on the AFLW2000 test set across training epochs in Fig. 7. It demonstrates that, as early as epoch five, the 6D-based model with Geodesic loss surpasses all other models across all epochs, showcasing its rapid convergence rate. This finding is coherent with our strong results for strong pitch and roll rotation in Fig 6.

TABLE VII
COMPARISON OF THE MAE BETWEEN THE DIFFERENT BACKBONES

	AFLW2000				BIWI			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
ResNet18	3.18	4.81	3.26	3.75	3.09	5.94	2.93	3.99
ResNet50	3.27	4.58	2.98	3.61	3.23	5.32	2.78	3.78

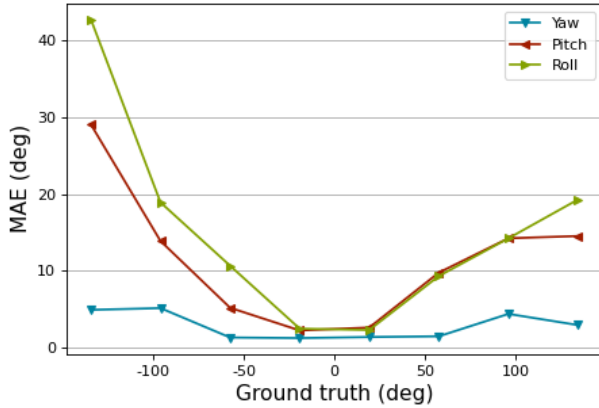


Fig. 8. Euler angle error for the CMU Panoptic + 300W-LP test set.

3) *Backbone*: In a final experiment, we analyze the impact of the chosen backbone on the results. Our results from Table I already proved the superiority of our 6D rotation matrix approach over other methods using the same backbone. Nevertheless, we want to evaluate the impact of the number of parameters on our results. In Table VII we compare our previous results with a model that was trained with the smaller ResNet18. It is remarkable, that our model that was trained on the 50% smaller backbone ResNet18 still achieves better results on the AFLW2000 dataset than all other methods from Table IV except one. For the BIWI dataset, the accuracy compared to ResNet50 is reduced only by a very small margin. This confirms that our model’s overall performance is predominantly accounted by our 6D rotation representation and hardly by the used backbone. Moreover, it shows that the commonly used ResNet50 is not necessary for achieving proper accuracy, as the more efficient ResNet18 reports similar performance. This becomes an important aspect, when the head pose estimation is used in settings with limited computational resources.

F. Limitations

Our model achieves accurate and robust prediction for an extended range of rotation. This especially applies for the yaw angle, which encounters the strongest rotations in common application scenarios. However, the roll and pitch can also reach strong rotations, that are only marginally represented in our training data (see Fig. 4). This can lead to reduced robustness and accuracy in application scenarios with unusual camera angles and head poses. To analyze this, we degreewise calculated the error of our 6DRepNet360 model on the test set of our CMU Panoptic + 300W-LP 70/30 split from section IV-D.2. The results are shown in Fig. 8 and illustrate that the error rate for the yaw angle is consistently low, while

the roll and pitch error rate increase with stronger rotations. This demonstrates that there is still a lack of training and also test data for this extended range of rotations. In our test set, only three samples exceed $[-100, +100]$ degrees in roll and only five samples exceed $[-100^\circ, +100^\circ]$ in pitch. In our experiments, we approached this limitation by performing image rotation augmentation that synthetically expands the roll and pitch range. Further, the CMU Panoptic dataset is taken in laboratory settings with similar background and lightning conditions. Additional data with stronger variation could therefore benefit the generalization performance as well.

V. CONCLUSION

In this paper, we tackle the major challenge of unconstrained full rotation head pose estimation that is a rarely explored research subject yet. First, we formulate a continuous 6D rotation matrix representation for an unambiguous and continuous appearance parameterization. This approach forms the basis for a stable and precise network training that we further optimize by introducing a geodesic distance based loss. With the use of the CMU Panoptic dataset, we accumulate a more comprehensive head pose dataset that exceeds the common public dataset in variety and size and allows us to create a model that is able to predict full head pose rotations. We evaluate our approach in multiple experiments that demonstrate that our 6D rotation representation achieves superior performance compared to the state-of-the-art and is able to efficiently learn the full range of head pose orientation. We complete our study with an ablation study to analyze the impact of the rotation representation, backbone and loss function on our results.

REFERENCES

- [1] A. Veronese, M. Racca, R. S. Pieters, and V. Kyrki, “Probabilistic mapping of human visual attention from head pose estimation,” *Frontiers Robot. AI*, vol. 4, p. 53, Oct. 2017.
- [2] S. O. Ba and J.-M. Odobez, “Recognizing visual focus of attention from head pose in natural meetings,” *IEEE Trans. Syst., Man, Cybern., B Cybern.*, vol. 39, no. 1, pp. 16–33, Feb. 2009.
- [3] T. Singh, M. Mohadikar, S. Gite, S. Patil, B. Pradhan, and A. Alamri, “Attention span prediction using head-pose estimation with deep neural networks,” *IEEE Access*, vol. 9, pp. 142632–142643, 2021.
- [4] F.-J. Chang, A. Tran, T. Hassner, I. Masi, R. Nevatia, and G. G. Medioni, “FacePoseNet: Making a case for landmark-free face alignment,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2017, pp. 1599–1608.
- [5] L. Wei and E.-J. Lee, “Multi-pose face recognition using head pose estimation and PCA approach,” *Int. J. Digit. Content Technol. Appl.*, vol. 4, no. 1, pp. 112–122, Feb. 2010.
- [6] A. Kumar, A. Alavi, and R. Chellappa, “KEPLER: Simultaneous estimation of keypoints and 3D pose of unconstrained faces in a unified framework by learning efficient H-CNN regressors,” *Image Vis. Comput.*, vol. 79, pp. 49–62, Nov. 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0262885618301549>
- [7] R. Ranjan, V. M. Patel, and R. Chellappa, “HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
- [8] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, “Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation,” in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2007, pp. 709–714.
- [9] S. Jha and C. Busso, “Estimation of driver’s gaze region from head position and orientation using probabilistic confidence regions,” *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 59–72, Jan. 2023.

- [10] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5494–5503.
- [11] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.
- [12] M. C. Bühler, A. Meka, G. Li, T. Beeler, and O. Hilliges, "VariTex: Variational neural face textures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13870–13879.
- [13] D. Strazdas, J. Hintz, and A. Al-Hamadi, "Robo-HUD: Interaction concept for contactless operation of industrial robotic systems," *Appl. Sci.*, vol. 11, no. 12, p. 5366, Jun. 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/12/5366>
- [14] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. P. de Ruiter, and A. Knoll, "Modelling state of interaction from head poses for social human-robot interaction," in *Proc. HRI*, 2012, pp. 1–6.
- [15] M. E. Foster, A. Gaschler, and M. Giuliani, "Automatically classifying user engagement for dynamic multi-party human-robot interaction," *Int. J. Social Robot.*, vol. 9, no. 5, pp. 659–674, Nov. 2017.
- [16] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, Art. no. 215509.
- [17] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.
- [18] 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, Jan. 2020, Art. no. 103827.
- [19] H. Hsu, T. Wu, S. Wan, W. H. Wong, and C. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.
- [20] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *Proc. AAAI*, 2020, pp. 1–8.
- [21] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang, "MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation," *IEEE Trans. Multimedia*, vol. 24, pp. 2449–2460, 2022.
- [22] N. Aghli and E. Ribeiro, "A data-driven approach to improve 3D head-pose estimation," in *Proc. Adv. Vis. Comput., 16th Int. Symp. (ISVC)*. Berlin, Heidelberg: Springer-Verlag, Oct. 2021, pp. 546–558, doi: [10.1007/978-3-030-90439-5_43](https://doi.org/10.1007/978-3-030-90439-5_43).
- [23] D. Dai, W. Wong, and Z. Chen, "Rankpose: Learning generalised feature with rank supervision for head pose estimation," in *Proc. 31st Brit. Mach. Vis. Conf. (BMVC)*, U.K.: BMVA Press, Sep. 2020, pp. 1–12. [Online]. Available: <https://www.bmvc2020-conference.com/assets/papers/0401.pdf>
- [24] Y. Yamaura, Y. Tsuboshita, and T. Onishi, "Head pose estimation for an omnidirectional camera using a convolutional neural network," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2018, pp. 1–5.
- [25] S. S. Mukherjee, R. H. Baxter, and N. M. Robertson, "Instantaneous real-time head pose at a distance," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3471–3475, doi: [10.1109/ICIP.2015.7351449](https://doi.org/10.1109/ICIP.2015.7351449).
- [26] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1070–1083, Jun. 2016.
- [27] K. Khan, R. U. Khan, R. Leonardi, P. Migliorati, and S. Benini, "Head pose estimation: A survey of the last ten years," *Signal Process., Image Commun.*, vol. 99, Nov. 2021, Art. no. 116479. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0923596521002332>
- [28] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5738–5746.
- [29] Y. Zhou and J. Gregson, "WheNet: Real-time fine-grained estimation for wide range head pose," in *Proc. 31st Brit. Mach. Vis. Conf. (BMVC)*, U.K.: BMVA Press, 2020, pp. 1–13. [Online]. Available: <https://www.bmvc2020-conference.com/assets/papers/0907.pdf>
- [30] H. Joo et al., "Panoptic studio: A massively multiview system for social motion capture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3334–3342.
- [31] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.
- [32] 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)*, Oct. 2022, pp. 2496–2500.
- [33] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
- [34] P. Werner, F. Saxen, and A. Al-Hamadi, "Landmark based head pose estimation benchmark and method," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3909–3913.
- [35] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2014, pp. 1867–1874.
- [36] A. Gupta, K. Thakkar, V. Gandhi, and P. Narayanan, "Nose, eyes and ears: Head pose estimation by locating facial keypoints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1977–1981.
- [37] S. Li, C. Xu, and M. Xie, "A robust O(n) solution to the perspective-n-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1444–1450, Jul. 2012.
- [38] S. Ohayon and E. Rivlin, "Robust 3D head tracking using camera pose estimation," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, 2006, pp. 1063–1066.
- [39] S. Tulyakov, L. A. Jeni, J. F. Cohn, and N. Sebe, "Viewpoint-consistent 3D face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2250–2264, Sep. 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:22084543>
- [40] H. Zhang, Q. Li, and Z. Sun, "Adversarial learning semantic volume for 2D/3D face shape regression in the wild," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4526–4540, Sep. 2019.
- [41] L. Zeng et al., "3D-aware facial landmark detection via multiview consistent training on synthetic data," in *Proc. CVPR*, 2023, pp. 1–12.
- [42] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.
- [43] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3658–3666.
- [44] L. Torgo and J. Gama, "Regression by classification," in *Advances in Artificial Intelligence*, D. L. Borges and C. A. A. Kaestner, Eds. Berlin, Germany: Springer, 1996, pp. 51–60.
- [45] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, K. Chaudhuri and R. Salakhutdinov, Eds. vol. 97, Jun. 2019, pp. 6105–6114. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [46] 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. (WACV)*, Jan. 2021, pp. 1188–1197.
- [47] Z. Liu, Z. Chen, J. Bai, S. Li, and S. Lian, "Facial pose estimation by deep learning from label distributions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 1232–1240.
- [48] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2874–2881, Aug. 2021.
- [49] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "Img2Pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7613–7623.
- [50] H. N. Viet, L. N. Viet, T. N. Dinh, D. T. Minh, and L. T. Quac, "Simultaneous face detection and 360 degree head pose estimation," in *Proc. 13th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2021, pp. 1–7.
- [51] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.
- [52] 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, Feb. 2013.
- [53] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

- [54] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge," in *Proc. ECCV Workshops*, 2016, pp. 616–624.
- [55] V. Le, J. Brandt, Z. L. Lin, L. D. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. ECCV*, 2012, pp. 679–692.
- [56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 896–903.
- [57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [58] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
- [61] A. Buslaev et al., "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020. [Online]. Available: <https://www.mdpi.com/2078-2489/11/2/125>



Thorsten Hempel was born in Pinneberg, Schleswig-Holstein, Germany, in 1993. He received the M.S. degree in industrial engineering from Otto von Guericke University Magdeburg, Germany, in 2019. Since 2019, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto von Guericke University Magdeburg. His research interests include computer vision, cognitive robotics, and human–robot interaction.



Ahmed A. Abdelrahman was born in Cairo, Egypt, in 1989. He received the B.Sc. and M.Sc. degrees in electrical and computer engineering from MTC. He is currently pursuing the Ph.D. degree in electrical engineering with Otto von Guericke University Magdeburg. Since 2021, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto von Guericke University Magdeburg. His research interests include computer vision, deep learning, and human–machine interaction.



Ayoub Al-Hamadi received the Ph.D. degree in technical computer science in 2001 and the Habilitation degree in artificial intelligence and the Venia Legendi degree in pattern recognition and image processing from Otto von Guericke University Magdeburg, Germany, in 2010. He is currently an Adjunct Professor and the Head of the Neuro-Information Technology Group, Otto von Guericke University Magdeburg. He is the author of more than 380 papers in peer-reviewed international journals, conferences, and books. His research interests include computer vision, pattern recognition, and image processing.