WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose

Yijun Zhou   Huawei Technologies Canada
yijun.zhou1@huawei.com

James Gregson
james.gregson@huawei.com

Abstract

We present an end-to-end head-pose estimation network designed to predict Euler angles through the full range head yaws from a single RGB image. Existing methods perform well for frontal views but few target head pose from all viewpoints. This has applications in autonomous driving and retail. Our network builds on multi-loss approaches with changes to loss functions and training strategies adapted to wide range estimation. Additionally, we extract ground truth labelling of anterior views from a current panoptic dataset for the first time. The resulting Wide Headpose Estimation Network (WHENet) is the first fine-grained modern method applicable to the full-range of head yaws (hence wide) yet also meets or beats state-of-the-art methods for frontal head pose estimation. Our network is compact and efficient for mobile devices and applications. Code will be available at: https://github.com/Ascend-Research/HeadPoseEstimation-WHENet

1 Introduction

Head pose estimation (HPE) is the task of estimating the orientation of heads from images or video (see Figure 2(a)) and has seen considerable research. Applications of HPE are wide-ranging and include (but are not limited to) virtual & augmented-reality [20], driver assistance, markerless motion capture [8] or as an integral component of gaze-estimation [19] since gaze and head pose are tightly linked [17].

The importance of HPE is well described in [19]. They describe wide-ranging social interactions such as *mutual gaze*, *driver-pedestrian* & *driver-driver* interaction. It is also important in providing visual cues for the targets of conversation, to indicate appropriate times for speaker/listener role switches as well as to indicate agreement [19]. For systems to interact with people naturally, it is important to be sensitive to head pose.

Most HPE methods target frontal to profile poses since applications are plentiful, the face is feature-rich and training datasets are widely available. However, covering full range is useful in many areas including driver assistance [21], motion-capture and to generate attention maps for advertising and retail [5, 1].

Furthermore, though methods such as [27, 29] perform well, they may be prohibitively large networks for mobile and embedded platforms. This is especially important in autonomous driving where many subsystems must operate concurrently.
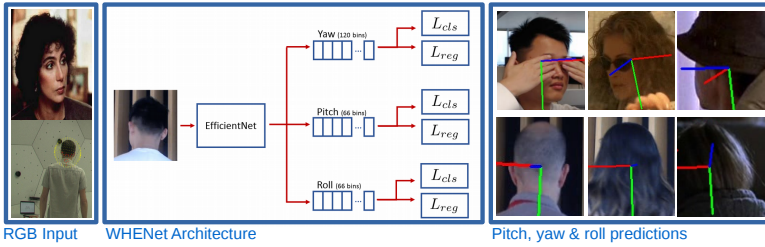
Figure 1: The WHENet full-range head pose estimation network (center) refines [27] by combining an EfficientNet[34] backbone with classification and regression losses on each of pitch, yaw and roll. A new *wrapped-loss* stabilizes the network at large yaws. Training this network using anterior views repurposed from [14] (lower-left) via a novel reprocessing procedure allows WHENet to predict head poses from anterior view as well as heads featuring heavy occlusions, fashion accessories and adverse lighting (right) with a mobile-friendly model. In spite of this, WHENet is as accurate or more accurate than existing state-of-the-art methods that are constrained to frontal or profile views. Some images are from [44]&[45]

Based on these criteria, we developed the Wide Headpose Estimation Network (WHENet) which extends head-pose estimation to the full range of yaws (hence *wide*) using a mobile-friendly architecture. In doing so, we make the following contributions:

- We introduce a *wrapped loss* that significantly improves yaw accuracy for anterior views in full-range HPE.

- We detail an automated labeling process for the CMU Panoptic Dataset [14] allowing it to be used for training and validation data in full-range HPE.

- We extend our modified network (WHENet) to the full range of yaws where it achieves state-of-the-art performance for full-range HPE and within 1.8% of state-of-the-art for narrow range HPE despite being trained for another task.

- We demonstrate that simple modifications to HopeNet [27] can achieve 29% improvement over the original network and $5-13\%$ improvement over the current state-of-the-art FSANet [41] for narrow range HPE from RGB images in the AFLW2000 [44] and BIWI [10] datasets.

2   Background & Related Work

Head pose estimation has been actively researched over the past 25 years. Approaches up to 2008 are well surveyed in [19] and much of our discussion on early methods follows their work. Due to the breadth of research activity and our target application of HPE from monocular RGB images, we exclude multi-view or active sensing methods from this review.

Classical methods include template matching and cascaded detectors. Template matching compares input images with a set of labeled templates and assign a pose based on nearby matches [22, 30]. They can have difficulty differentiating between identity and pose similarity [19]. Cascaded detectors train distinct detectors that also localize heads for each

(discretized) pose. Challenges include prediction resolution and resolving when multiple detectors fire [19].

Geometric & deformable models are similar methodologies. Geometric models use features, e.g. facial keypoints, from the input images and analytically determine the matching head pose using a static template [5]. The majority of their complexity lies in detecting the features, which is itself a well studied problem, see e.g. [12, 53] and survey [56]. Deformable models are similar but allow the template to deform to match subject-specific head features. Pose or other information is obtained from the deformed model [6, 39, 42, 44].

Regression & classification methods serve as supersets or components of most other methods. Regression methods use or fit a mathematical model to directly predict pose based on labeled training data. The formulation of the regressor is wide-ranging and includes principle component analysis [52, 45] and neural networks [18, 27] among others. In contrast, classification methods predict pose from a discretized set of poses. Prediction resolution tends to be lower, generally less than 10-12 discrete poses. Methods include decision trees & random forests [2, 9], multi-task learning [37, 38] and neural networks [18, 27]. Our networks are most similar to the multi-loss framework in [27] which uses classification and regression objectives. Other works employ soft stage-wise regression [40, 41] by training with both classification and regression objectives at multiple scales.

Multi-tasks methods combines the head pose estimation problem with other facial analysis problems. Studies[7, 23, 24, 43] show that learning related tasks at the same time can achieve better performance than training tasks individually. For instance, in [23, 24] face detection and pose regression are trained jointly.

Full-range methods are much less common than narrow-range since most existing datasets for HPE focus on frontal to profile views. Recent methods include [11, 25, 26] which classify poses into coarsely-grained bins/classes to determine yaw. Unlike our method, pitch and roll are not predicted in [11, 25, 26].

Datasets for facial pose include BIWI [10], AFLW2000[44] and 300W-LP [44]. Both AFLW2000 and 300W-LP use a morphable model fit to faces under large pose variation and report Euler angles. 300W-LP generates additional synthetic views to enlarge the dataset. More recently, the UMD Faces[1] and Pandora[3] datasets provide a range of data labels, including head pose. A disadvantage for our application is that they do not cover the full-range of head poses but only frontal-to-profile views.

Very important to our method is the CMU Panoptic Dataset [14]. It captures subjects from an abundance of calibrated cameras covering a full-hemisphere and provides facial landmarks in 3D. Using this data, we are able to estimate head pose from near-frontal views and use this pose to label non-frontal viewpoints. By doing so, we cover the full range of camera-relative poses, allowing our method to be trained with anterior views. This is discussed further in Section 4

3 Our Method

Our network design is derived from the multi-loss framework of Ruiz et al. [27]. They combine a convolutional backbone with separate fully-connected networks that classify each of pitch, yaw and roll into $3°$ bins using softmax with a cross-entropy loss. A mean squared error (MSE) regression loss is also applied between the ground-truth labels and the expected value of the softmax output. The two losses are weighted to produce the final training objective for each angle.
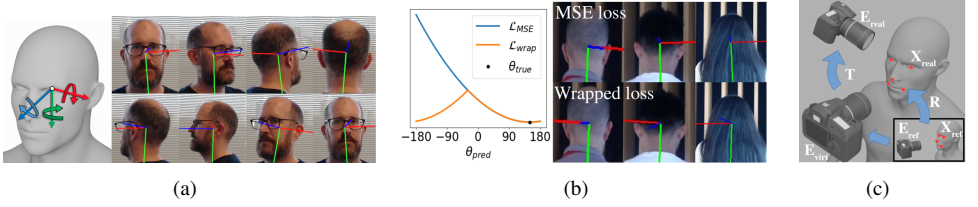
Figure 2: Head pose (a) is parameterized by pitch (red-axis), yaw (green-axis) and roll (blue-axis) angles in indicated directions. Our proposed wrapped loss function (b-left) avoids over-penalizing predictions $> 180°$ from true where pose is similar but MSE produces extreme loss values. This improves predictions when subjects face away from camera (b-right) where networks trained with MSE have errors approaching $180°$ for yaw. Note the x (red) axis should align with the left ear of subjects. In (c), we compute virtual camera extrinsics oriented to provide a frontal view along with true extrinsics to extract Euler angles from the CMU Panoptic dataset [14]. This allows us to automatically label tens of thousands of anterior view images for the full-range HPE task. We believe we are the first to do so.

In [27], Ruiz *et al*. suggest that this combination of losses gives the resulting network robustness & stability due to the softmax & cross-entropy loss while still providing fine-grained supervision and output via the regression loss.

We adopt this overall framework from [27] but make substitutions for both loss functions in order to adapt the method to full-range. To our knowledge we are the first to do so. Yaw prediction is divided into 120 $3°$ bins covering the full range of yaws $(-180°, 180°]$. Pitch and roll predictions are each made from 66 $3°$ bins covering the range $[-99°, 99°]$, although only the bins from $[-90, 90]$ are ultimately used. The regression and classification losses for each of pitch, yaw and roll are combined via:

$$\mathcal{L} = \alpha \mathcal{L}_{reg} + \beta \mathcal{L}_{cls} \tag{1}$$

Here $\mathcal{L}_{cls}$ is the classification loss and $\mathcal{L}_{reg}$ is the regression loss while α and β trade off the influence of one with respect to the other.

We tested different classification losses and chose a sigmoid activation with binary cross-entropy for $\mathcal{L}_{cls}$. This differs from [27] and shows marginally improved accuracy for the full-range task but is mentioned primarily for reproducibility.

As in [27] we predict output angles from the bin logits by applying softmax to obtain bin probabilities and take an expectation of the result for each of yaw, pitch and roll:

$$\theta_{pred} = 3 \sum_{i=1}^{N} p_i \left(i - \frac{1+N}{2} \right) \tag{2}$$

Here p_i is the probability of the i'th bin, 3 is the bin width in degrees and N is the bin-count of either 120 (yaw) or 66 (pitch and roll). The subtracted term shifts bin indices to bin centres.

In [27], a mean squared error (MSE) loss was used for regression which was sufficient for the limited range of yaws that were targeted. However, on the full-range task, this loss leads to erratic behaviour for subjects with absolute yaws exceeding $150°$. This is illustrated in Figure 2(b) and is due to $\pm180°$ having wildly different angles for the same pose.

To prevent this, we define a *wrapped loss* (see Figure 2(b)) that avoids this behavior. Rather than penalizing angle directly, it penalizes the minimal rotation angle that is needed

to align each yaw prediction with its corresponding dataset annotation:

$$\mathcal{L}_{wrap}(\theta_{pred}, \theta_{true}) = \frac{1}{N_{batch}} \sum_{i}^{N_{batch}} \min[|\theta_{pred}^{(i)} - \theta_{true}^{(i)}|^2, (360 - |\theta_{pred}^{(i)} - \theta_{true}^{(i)}|)^2] \quad (3)$$

The loss function $\mathcal{L}_{wrap}$ is plotted in Figure 2(b) for the target value $\theta_{true} = 150°$ and compared with an MSE loss. In the range $(-30, 180]$ the two are identical but diverge for angles $< -30°$. As they diverge MSE increases rapidly but the wrapped loss decreases as poses become more similar. The wrapped loss is smooth & differentiable everywhere except at a cusp $180°$ from the ground-truth yaw. This does make training the network more difficult, we suspect this is due to the cusp occurring at maximal errors in yaw.

The wrapped loss is key to the method's performance in anterior views, demonstrated empirically in Figure 2(b). A MSE trained network predicts entirely incorrect head poses while our predictions are consistent with the images. Figure 3 plots errors as a function of angles and shows average errors at large yaw are decreased by more than 50%.

In [23, 27] AlexNet and ResNet50 were used as backbones. These are quite large networks for a highly specialized task such as HPE. Since a focus of our method is to be mobile-friendly, we instead opted for a lighter backbone: EfficientNet-B0[34]. EfficientNet-B0 is the baseline model of the EfficientNet family that incorporates Inverted Residual Blocks (from MobileNetV2[28]) to reduce the number of parameters while adding skip connections.

This model size reduction is important on low-power embedded devices where head pose estimation may be only one component of a larger system. We have successfully ported a preliminary implementation to a low-power embedded platform where we see inference speeds approaching 60fps

4 Datasets & Training

Currently, major datasets for HPE are 300W-LP[44], AFLW2000[44] and BIWI[10]. Both AFLW2000 & 300W-LP use 3D Dense Face Alignment (3DDFA) to fit a morphable 3D model to 2D input images and provide accurate head poses as ground-truth. 300W-LP additionally generates synthetic views, greatly expanding the number of images.

The BIWI dataset is composed of video sequences collected by a Kinect. Subjects moved their heads trying to span all possible angles observed from a frontal position. Pose annotations were created from the depth information.

We follow the convention of [27] by reserving AFLW2000 [44] and BIWI [10] for testing, while using 300W-LP [44] for training.

Unfortunately, none of datasets mentioned above provide examples with (absolute) yaws larger than $100°$. To overcome this, we generated a new dataset combining 300W-LP with data from the CMU Panoptic Dataset [14]. The CMU Panoptic Dataset captured video of subjects performing tasks in a dome from approximately 30 HD cameras.

The panoptic dataset includes 3D facial landmarks and calibrated camera extrinsics and intrinsics but does not include head pose information. We use the landmarks and camera calibrations to locate and crop images of subjects' heads and to compute the corresponding camera-relative head pose Euler angles. We believe we are the first to use this dataset in this context.

Figure 1 shows a frame from the dataset. The panoptic dataset has very little background variation and so cannot be used to train networks alone since networks do not learn features

to differentiate subjects from general backgrounds. This is the motivation to combine it with 300W-LP, which is used for yaws in $(-99°, 99°)$ while the panoptic dataset provides data mostly outside this range.

Processing the CMU Panoptic Dataset. To compute camera-relative head pose Euler angles from the panoptic dataset we use the following procedure, depicted graphically in Figure 2(c). We first define a set of reference 3D facial landmarks (x_{ref}) matching the panoptic dataset facial keypoints annotations using a generic head model, except for the noisy jawline keypoints. A reference camera with intrinsics & extrinsics (K_{ref},E_{ref}) is then positioned to obtain a perfectly frontal view (yaw=pitch=roll=0) of x_{ref}. For each subject in each frame, we estimate the rigid transformation R between x_{ref} and the true keypoints x_{real} provided by the panoptic dataset using [13]. We then construct new extrinsics for a virtual camera $E_{virt} = E_{ref}R^{-1}$. This provides a (nominally) frontal view of the subject as they are positioned within the dome. Using the known real camera extrinsics E_{real} from the panoptic dataset and E_{virt}, we recover the rigid transformation $T = E_{real}E_{virt}^{-1}$ between the virtual camera extrinsics and each real camera. Finally we extract Euler angles in pitch-yaw-roll (x-y-z) order from T, mirroring yaw and roll in order to match the rotation convention from the datasets. There is a translation component to the rigid transform but it is not needed to determine the head orientation with respect to the camera optical axes.

Using the synthetic reference and virtual camera in this process ensures that we have a consistent and automatic method for labeling. It also considerably reduces noise compared to manually selecting a frontal camera and annotating corresponding Euler angles.

To crop images, we define a spherical set of points around the subject's head as a *helmet* and project these into each view to determine a cropping region. Inspired by Ming et al.[29], we leave a margin from the head bounding box so the network can learn to distinguish foreground and background. We followed the method of [29], using K = 0.5 for data from 300W-LP and adjust the helmet radius to 21cm for CMU data.

Training of WHENet is devided into two stages. We first train a narrow range model, WHENet-V with prediction range between $[-99°, 99°]$ for yaw, pitch and roll. We train this narrow range network on the 300W-LP [44] dataset using an ADAM optimizer with learning rate 1e-5. Full-range WHENet with 120 yaw bins is then trained starting from the WHENet-V weights using our full-rage data set combining 300W-LP and CMU Panoptic Dataset data. We use the same optimizer and learning rate for this step. This two-stage approach helps the full-range network converge better and learn more useful features since the CMU data has little background variation. During training, images are randomly downsampled by up to 15X to improve robustness.

Training data from 300W-LP and the panoptic dataset is used in nearly equal amounts, with the former providing narrow range samples and the latter primarily handling wide range. A small amount of panoptic data is also used to level a dip in the histogram of narrow range images near yaw=0.

5 Results & Discussion

For wide range results in this section, we define the Absolute Wrapped Error (AWE) as $AWE = \min(|\theta_{pred} - \theta_{true}|, 360 - |\theta_{pred} - \theta_{true}|)$ which properly handles wrapping of yaws. We also define Mean AWE (MAWE) as the arithmetic mean of AWE. For results below, we are using $\alpha = 1$, $\beta = 1$ for WHENet and $\alpha = 0.5$, $\beta = 2$ for WHENet-V based on our ablations. Hyperparameter details can be found in our supplementary.

Table 1 summarizes key results from WHENet and WHENet-V. We compare with eight narrow & two full-range methods and report, where applicable/available, the output range, number of parameters, mean average errors on BIWI [10], AFLW-2000 [44], an average of both (to indicate generalization) and reported full-range MAE if applicable. WHENet & WHENet-V are trained on 300W-LP [44] or our combined dataset and are not trained on the AFLW2000 or BIWI datasets, as is standard practice. We begin with a discussion of the full-range network WHENet, as this is our primary application.

Table 1: Summary of results: 'Aggr. MAE': average of BIWI and AFLW2000 overall results. 'Full MWAE': full-range results. **first/only**, **second**, -: not reported

Method	Full Range	Params $(\times 10^6)$	BIWI MAE°	AFLW2k MAE°	Avg. MAE°	Full MAWE°
KEPLER [16]	N	-	13.852	-	-	-
Dlib (68 points) [15]	N	-	12.249	15.777	14.013	-
FAN (12 points) [4]	N	6-24	7.882	9.116	8.499	-
3DDFA [44]	N	-	19.07	7.393	13.231	-
Shao et al. (K=0.5) [29]	N	24.6	5.999	5.478	5.875	-
Hopenet [27]	N	23.9	4.895	6.155	5.525	-
SSR-Net-MD [40]	N	**0.2**	4.650	6.010	5.330	-
FSA-Caps-Fusion [41]	N	**1.2**	4.000	**5.070**	**4.535**	-
Rehder et al. [26]	Yaw	-	-	-	-	19.0
Raza et al. [25]	Yaw	-	-	-	-	**11.25**
WHENet-V	N	4.4	**3.475**	**4.834**	**4.155**	-
WHENet	Y	4.4	**3.814**	5.424	4.619	**7.655**

Full-range WHENet results and comparisons are shown in Table 1. Of the three full-range methods compared, WHENet is the only method to also predict pitch and roll. Comparisons are difficult to perform objectively since both Raza et al. [25] and Rehder et al. [26] predict yaws only (no pitch and roll) and do so on non-public datasets. For [25], MAE was not reported, so we report the lowest possible errors given a uniform distribution of yaws with their bin-widths (i.e. the result if their method performed perfectly). WHENet still shows a 31% improvements, although these results should be treated as qualitative given lack of consistent testing data. This latter point is something we attempt to address in this work by making our dataset processing code for the CMU Panoptic Dataset [14] public.

Additionally, excluding WHENet-V, full-range WHENet has the lowest average errors on BIWI, second-lowest errors on AFLW2000 and has second lowest average overall errors, missing first place by 0.084° (1.8%) to FSANet [41]. This is significant since full-range WHENet was not trained specifically for the narrow range task, yet is still state-of-the-art or very competitve with FSANet [41], a significantly more complex network. In handling full-range inputs, WHENet is significantly more capable than FSANet, yet sacrifices remarkably little accuracy.

Figure 2(a) shows pose predictions generated using WHENet to track a subject rotating through a full revolution of yaw. WHENet produces coherent predictions throughout, even when the face is completely absent. A key to achieving this is our wrapped loss function, which significantly improves anterior views. Without the wrapped loss predictions are erratic and often nearly 180° from the true pose as shown in Figure 2(b). This improvement is enabled by our annotation process for the CMU panoptic dataset [14] which allows us to automatically generate labelings for tens of thousands of anterior views to tackle this chal-

Table 2: Comparison with state-of-the-art on BIWI[10](left) and AFLW2000[44](right) dataset. Best (bold underlined) and second best (bold) results are highlighted.

Method	BIWI				AFLW2000			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
KEPLER [16]	8.80	17.3	16.2	13.9	-	-	-	-
Dlib (68 points) [15]	16.8	13.8	6.19	12.2	23.1	13.6	10.5	15.8
FAN (12 points) [4]	-	-	-	-	8.53	7.48	7.63	7.88
3DDFA [44]	36.20	12.30	8.78	19.10	5.40	8.53	8.25	7.39
Hopenet(a = 1) [27]	4.81	6.61	3.27	4.90	6.92	6.64	5.67	6.41
Hopenet(a = 2) [27]	5.12	6.98	3.39	5.12	6.47	6.56	5.44	6.16
Shao *et al.* (K=0.5) [29]	4.59	7.25	6.15	6.00	5.07	6.37	4.99	5.48
SSR-Net-MD [40]	4.49	6.31	3.61	4.65	5.14	7.09	5.89	6.01
FSA-Caps-Fusion [41]	4.27	4.96	**2.76**	4.00	**4.50**	**6.08**	**4.64**	**5.07**
WHENet-V	**_3.60_**	**_4.10_**	**_2.73_**	**_3.48_**	**_4.44_**	**_5.75_**	**_4.31_**	**_4.83_**
WHENet	**3.99**	**4.39**	3.06	**3.81**	5.11	6.24	4.92	5.42

lenge.

Narrow range results and comparisons are listed in Tables 2. Dlib[15], KEPLER[16] and FAN[4] are all landmark based methods. 3DDFA[44] uses CNN to fit a 3D dense model to a RGB image. Hopenet[27] and FSANet[41] are two landmark-free methods both of which explore the possbility of treating a continuous problem (head pose) into different classes/stages. As showed in Tables 2, WHENet-V achieves state-of-the-art accuracy on both BIWI [10] and AFLW2000 [44], and outperforms the previous state-of-the-art FSANet [41] by 0.52° (13.1%) and 0.24° (4.7%) respectively, leading to an aggregate improvement of 0.38° (8.4%) on the two datasets overall. It also achieve first place in every metric. Interestingly, WHENet-V shows a 29% and 21% improvement over HopeNet [27] on these datasets, despite of having a very similar network architecture.

Figure 3: Histogram of the errors on our hybrid dataset for WHENet and WHENet (MSE), errors at high yaw are signficantly reduced with our proposed wrapped loss.
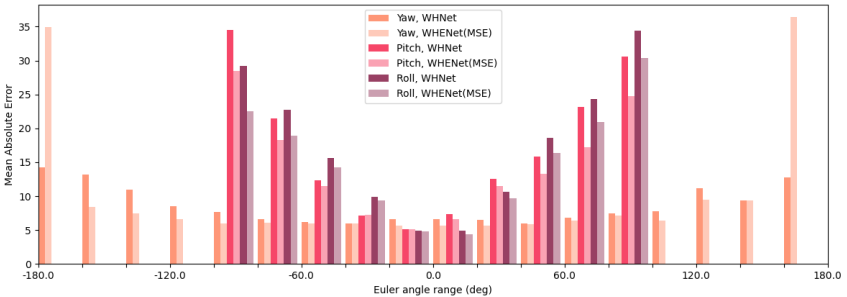
Ablation studies for the change of backbone and loss functions are shown in Figure 3 and Table 3. Figure 3 plots the mean errors at different angle intervals between wrapped regression loss (WHENet) and MSE loss (WHENet MSE).The mean error in yaw for extreme poses (close to -180/180 degrees) reaches 35 degrees when using MSE loss. By introducing the wrapped loss, errors for these angles are reduced by more than 50% and are more consistent with those from medium or low yaws. High pitch and roll errors are typical in HPE methods as yaws approach $\pm 90°$ as seen in [29], we believe due to gimbal lock in the data

labeling. Additionally, a quantitative result comparison can be found in Table 3. WHENet CE uses cross-entropy loss for classification and wrapped regression loss for regression. WHENet MSE uses the binary cross-entropy loss for classification and MSE loss for regression. WHENet uses binary cross-entropy loss for classification and wrapped regression loss for regression. We can see in the wide-range dataset (combine), WHENet has superior performance in yaw which confirms our observations in Figure 3. Although WHENet sacrifices some performance in the narrow range testing compared with other two loss settings, our main focus is in the wide-range predictions where we see significantly improved errors at large yaws. We believe performance could be improved by adapting the template keypoints used in annotating the CMU dataset to each subject rather than using a fixed template but leave this to future work.

Table 3: Result comparison of different loss settings. Combine indicates our combined dataset of CMU panoptic and 300W-LP

	Combine				BIWI				AFLW2000			
	Yaw	Pitch	Roll	MWAE	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
WHENet CE	8.75	7.65	6.74	7.71	**3.51**	**4.13**	**3.04**	**3.56**	5.50	6.36	4.94	5.60
WHENet MSE	9.69	**7.15**	**6.19**	7.68	3.79	4.82	3.21	3.94	**5.03**	6.41	5.16	5.53
WHENet	**8.51**	7.67	6.78	**7.66**	3.99	4.39	3.06	3.81	5.11	**6.24**	**4.92**	**5.42**

The supplement provides additional ablation studies on the metaparameters α and β as well the effects of resolution and relative comparison to video and depth-based methods that are outside our target application.

6 Conclusions and future work

In this paper we have presented WHENet, a new method for HPE that can estimate head poses in the full 360 degree range of yaws. This is achieved by careful choice of our wrapped loss function as well as by developing an automated labeling method for the CMU Panoptic Dataset [14]. We believe we are the first to adapt this dataset to the specific task of head-pose estimation. WHENet meets or exceeds the performance of state-of-the-art methods tuned for the specific task of frontal-to-profile HPE in spite of being trained for the full range of yaws. We are not aware of competing methods with similar capabilities and accuracy.

In the future, we would like to extend upon this work by reducing network size even further. Reducing input image resolution has the potential to lower memory usage as well as allow shallower networks with fewer features. This could yield even further improvements in speed and size of network.

Another interesting avenue is modifying the representation of head pose. Euler angles are minimal and interpretable but have the drawback of gimbal lock. The pitch-yaw-roll rotation ordering of existing datasets such as AFLW2000 and BIWI emphasize this effect at yaws near $\pm 90°$ where pitch and roll effectively counteract each other. We believe this leads to the relatively high pitch and roll errors near profile poses seen in Figure 3 which can also be seen in methods such as Shao et al. [29]. Relabeling data to yaw-pitch-roll order might reduce this. Similarly other rotation representations such as axis-angle, exponential map or quaternions may help, though adapting the architecture to alternative rotation representations is likely non-trivial due to the classification networks.

7 Acknowledgment

References

[1] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[2] Ben Benfold and Ian D Reid. Colour invariant head pose classification in low resolution video. In *BMVC*, pages 1–10, 2008.

[3] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5494–5503. IEEE, 2017.

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[5] Patrick Burger, Martin Rothbucher, and Klaus Diepold. Self-initializing head pose estimation with a 2d monocular usb camera. Technical report, Lehrstuhl für Datenverarbeitung, 2014.

[6] Qin Cai, David Gallup, Cha Zhang, and Zhengyou Zhang. 3d deformable face tracking with a commodity depth camera. In *European conference on computer vision*, pages 229–242. Springer, 2010.

[7] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European conference on computer vision*, pages 109–122. Springer, 2014.

[8] Márcio Cerqueira de Farias Macedo, Antônio Lopes Apolinário, and Antonio Carlos dos Santos Souza. A robust real-time face tracking using head pose estimation for a markerless ar system. In *2013 XV Symposium on Virtual and Augmented Reality*, pages 224–227. IEEE, 2013.

[9] Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. In *CVPR 2011*, pages 617–624. IEEE, 2011.

[10] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458, February 2013.

[11] DuYeong Heo, Jae Yeal Nam, and Byoung Chul Ko. Estimation of pedestrian pose orientation using soft target training based on teacher–student framework. *Sensors*, 19 (5):1147, 2019.

[12] Rainer Herpers, Markus Michaelis, K-H Lichtenauer, and Gerald Sommer. Edge and keypoint detection in facial regions. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 212–217. IEEE, 1996.

[13] Berthold KP Horn, Hugh M Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A*, 5(7):1127–1135, 1988.

[14] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):190–204, 2017.

[15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[16] Amit Kumar, Azadeh Alavi, and Rama Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 258–265. IEEE, 2017.

[17] Stephen RH Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5):752–771, 2004.

[18] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

[19] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2008.

[20] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *IEEE Transactions on intelligent transportation systems*, 11(2):300–311, 2010.

[21] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 709–714. IEEE, 2007.

[22] Jeffrey Ng and Shaogang Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image and Vision Computing*, 20(5-6):359–368, 2002.

[23] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2017.

[24] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[25] Mudassar Raza, Zonghai Chen, Saeed-Ur Rehman, Peng Wang, and Peng Bao. Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, 2018.

[26] Eike Rehder, Horst Kloeden, and Christoph Stiller. Head detection and orientation estimation for pedestrian safety. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2292–2297. IEEE, 2014.

[27] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018.

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[29] Mingzhen Shao, Zhun Sun, Mete Ozay, and Takayuki Okatani. Improving head pose estimation with a combined loss and bounding box margin adjustment. *arXiv preprint arXiv:1905.08609*, 2019.

[30] Jamie Sherrah, Shaogang Gong, and Eng-Jon Ong. Understanding pose discrimination in similarity space. In *BMVC*, pages 1–10, 1999.

[31] Teera Siriteerakul. Advance in head pose estimation from low resolution images: A review. *International Journal of Computer Science Issues*, 9(2):1, 2012.

[32] Sujith Srinivasan and Kim L Boyer. Head pose estimation using view based eigenspaces. In *Object recognition supported by user interaction for service robots*, volume 4, pages 302–305. IEEE, 2002.

[33] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[34] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[35] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. Context-aware cnns for person head detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2893–2901, 2015.

[36] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65, 2018.

[37] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *Proceedings of the IEEE international conference on computer vision*, pages 1177–1184, 2013.

[38] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083, 2015.

[39] Ruigang Yang and Zhengyou Zhang. Model-based head pose tracking with stereovision. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 255–260. IEEE, 2002.

[40] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, page 7, 2018.

[41] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019.

[42] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[43] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[44] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[45] Youding Zhu and Kikuo Fujimura. Head pose estimation for driver monitoring. In *IEEE Intelligent Vehicles Symposium, 2004*, pages 501–506. IEEE, 2004.