# ALL IN ONE NETWORK FOR DRIVER ATTENTION MONITORING

*Dawei Yang*[1], *Xinlei Li*[1], *Xiaotian Dai*[3], *Rui Zhang*[2], *Lizhe Qi*[1], *Wenqiang Zhang*[2*], *Zhe Jiang*[3*]

[1]Academy for Engineering and Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
[3]Department of Computer Science, University of York, UK

## ABSTRACT

Nowadays, driver drowsiness and driver distraction is considered as a major risk for fatal road accidents around the world. As a result, driver monitoring identifying is emerging as an essential function of automotive safety systems. Its basic features include head pose, gaze direction, yawning and eye state analysis. However, existing work has investigated algorithms to detect these tasks separately and was usually conducted under laboratory environments. To address this problem, we propose a multi-task learning CNN framework which simultaneously solve these tasks. The network is implemented by sharing common features and parameters of highly related tasks. Moreover, we propose Dual-Loss Block to decompose the pose estimation task into pose classification and coarse-to-fine regression and Objectcentric Aware Block to reduce orientation estimation errors. Thus, with such novel designs, our model not only achieves SOA results but also reduces the complexity of integrating into automotive safety systems. It runs at 10 fps on vehicle embedded systems which marks a momentous step for this field. More importantly, to facilitate other researchers, we publish our dataset FDUDrivers which contains 20000 images of 100 different drivers and covers various real driving environments. FDUDrivers might be the first comprehensive dataset regarding driver attention monitoring.

***Index Terms*—** Driver attention, Driver monitoring system, Drowsiness, Distraction, CNN

## 1. INTRODUCTION

According to World Health Organization [1], motor vehicle accidents are one of the leading causes of death in the world. Traffic accidents kill approximately 1.35 million people every year, with at least 50 million people suffering non-fatal or fatal injuries such as a disability. Road accidents occur for a variety of reasons. The lack of attention (drowsiness and distraction) in the driving situation is regarded as the major reason, accounting for 41% [2]. Although with the evolution of autonomous driving that is believed to bring better road-safety guarantee compared to manual driving [14], today's most advanced vehicles still only equipped with partial and conditional automation, requiring frequent driver action. The fact is that the monotony of such a scenario may induce fatigue or distraction, reducing driver awareness and impairing the regain of the vehicle's control. Therefore, driver attention monitor is emerging as an essential requirement for automotive safety systems, aiming to identify potential risks and prevent accidents happening.

Basic driver attention monitoring features include head pose, gaze direction, yawning and eye state analysis blink rate, blink duration, eye open/close. Each of these features can indicate whether the driver is distracted and drowsy to some extent. More recently, computer vision based approaches [3] [4] [5] [6] [7] [8] [9] are becoming popular, compared to physiological based methods that require biological sensors to be attached on the driver. This is because computer vision based methodologies only use a camera to monitor and analyze drivers behaviors without disturbing or annoying the driver. However, existing research has investigated algorithms to perform each task separately, e.g., one training model for face alignment, and another for head pose prediction. The first problem of them is that each task requires its own storage and memory to execute the model, significantly lifting computational cost that is unpopular to automotive safety systems. Secondly, integrating multiple different models leads the system end up very complex, error-prone, and less compatible. The situation will become worse if the input of some models depends on the output of other models, but the two are not compatible. Thirdly, especially in a CNN network, focusing on a single task increases the risk of overfitting. This is because some information coming from the training signals of related tasks that might help networks do even better on the metric is ignored. So far, there is no single solution that can address all the problems mentioned above. Besides algorithms, another main challenge in this field is the lack of image dataset for driver attention monitoring. Current datasets have been collected either under controlled laboratory conditions or under general life situations that are both far from real driving environments.

This paper introduces a new CNN framework that is specifically tailored for driver attention monitoring. Because the automotive

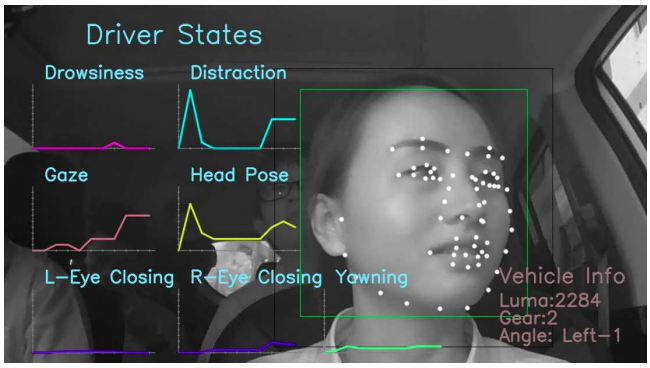


**Fig. 1**. A demonstration application of Driver Monitor System. The indicators of the driver states in the figure are derived from the output of our proposed DANet. For example, the model precisely locates the position of the face landmark and the yawning state can be estimated by tracking the distance between upper and lower lips. Finally, these time series indicators are combined to determine whether the drive is fatigue or distracted.

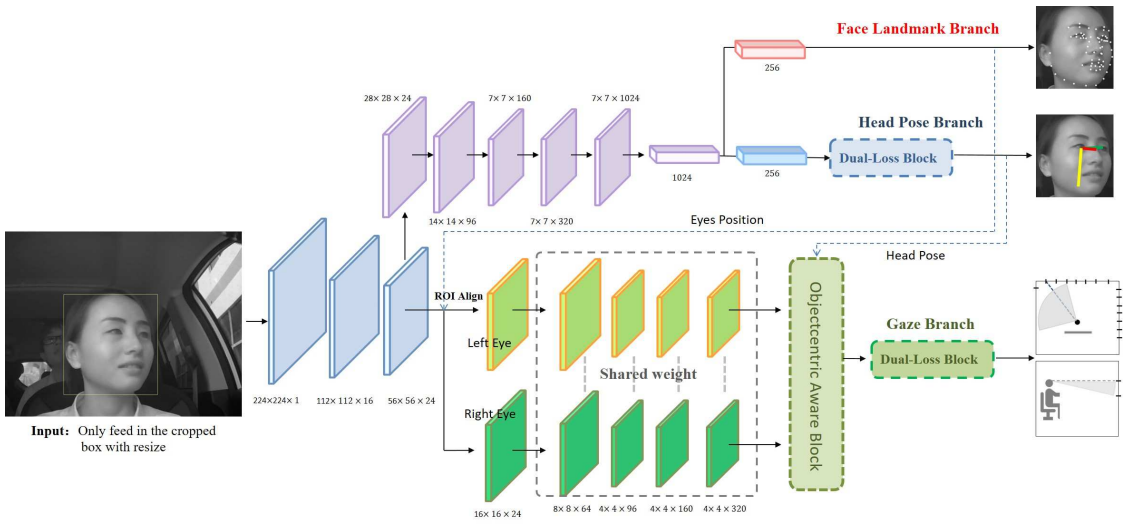


**Fig. 2**. The architecture of DANet which is a multi-task learning CNN framework with jointly parameter sharing. Given a driver face image, it can simultaneously output the drivers face landmark, head pose and gaze direction through three branches. Once the face landmark is detected, the rest states (yawning and eye state blink rate, blink duration, eye open/close.) can be easily computed.

safety systems has high requirements for accuracy and speed, we pursue the goal of developing a multi-task learning network that can simultaneously solve all the tasks and generalize better on each task for driver attention monitoring. The main contributions of our work are highlighted as follows:

1. Proposing a multi-task learning CNN framework (DANet) with jointly parameter sharing. This is the first work simultaneously solving a diverse set of driver face analysis tasks using a single CNN in an end-to-end manner.

2. Proposing Dual-Loss Block (DLB) to decompose the pose estimation task into pose classification and coarse-to-fine regression and Objectcentric Aware Block (OAB) to reduce orientation estimation errors and accelerate the training process.

3. Achieving state-of-art performances on each task, while it runs in 100ms for a 224x224 drivers face image on the vehicle embedded system (Quad ARM Cortex-A57). 100ms (10 fps) is sufficient for real-time monitoring and nearly 4 times faster than the sum of execution time of these separate models, reduced by 75%. The real-time performance marks a momentous step for driver attention monitoring.

4. Publishing the first comprehensive driver attention monitoring dataset - FDUDrivers https://github.com/FDUXilly/FDUDrivers. The dataset contains 20000 images of 100 different drivers and it was collected under various real driving environments. We believe that FDUDrivers will serve as an invaluable resource for researchers in this field, accelerating the development of driver attention monitoring.

## 2. METHOD

In this section, we describe the advantages of multi-task learning in the context of face analysis and provide the details of the network design. We explain how combined classification and regression can be used to improve performance and introduce our proposed Dual-loss block. Lastly, we provide a detailed analysis of associating with predicted head pose makes great contribution to gaze estimation.

### 2.1. Multi-task Learning

In deep learning, the key is to optimize a task for a particular metric. To achieve this, a single model or an ensemble of models is usually trained to perform the desired task. Although we can generally obtain acceptable performance this way, by being laser-focused on a single task, we ignore information that might help us do even better on the metric we care about. Specifically, this information comes from the training signals of related tasks. By sharing representations between related tasks, the model is able to generalize better on each original task. This approach is called Multi-Task Learning.

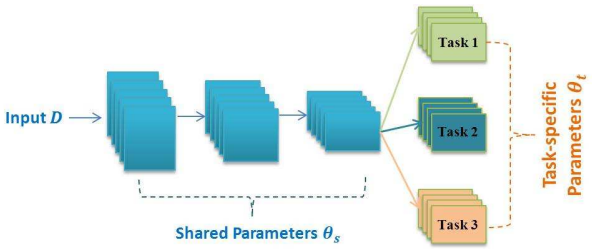


**Fig. 3**. Parameter sharing for multi-task learning

The motivation behind this is intuitive. Adopting multi-task learning forces the network to learn the correlations between data from different distributions in an effective way. The shadow layers learn general representation common to different yet highly correlated tasks, whereas upper layers are more specific to the given task. As a result, the network ends up regularizing the shared parameters which reduces the risk of overfitting and building a synergy among different domains and tasks. Rajeev Ranjan in [10] argues that the optimization goal for multi-task learning can be illustrated as the following equation:

$$\left(\theta_s^*, \theta_{t_i}^*\right) = \underset{\left(\theta_s, \theta_{t_i}\right)}{\arg\min} J_i\left(\theta_s, \theta_{t_i}; D\right) + \lambda R_i\left(\theta_s; D\right) \quad (1)$$

where $\lambda_i$ is a weight to balance the contribution of task $t_i$, and $R_i$ is

a regularizer on $\theta_s$ with respect to task $t_i$.

In DANet, as presented in Fig.2, the landmark branch and the head pose branch share most of the network parameters except the last two FC layers that includes task-specific parameters. We argue that this setting is reasonable because intuitively there is a direct correlation between head pose and the distribution of the landmarks, and we also verify the effectiveness of the setting in the evaluation, see Table 3. As for gaze estimation branch, it shares the low-level features from the first three layers, because it will require more high-level features to be extracted from a certain area where only accounts for a small part of landmark and head pose.

Each branch has its own task-specific loss function, which will be discussed in the next two sections. The overall loss function used to optimize DANet is the sum of these losses, as the following:

$$\mathcal{L} = \mathcal{L}_{ldmk} + \lambda_1 \cdot \mathcal{L}_{hp} + \lambda_2 \cdot \mathcal{L}_{gaze} \tag{2}$$

where $\lambda_1$ and $\lambda_2$ is the weight corresponding to head pose and gaze respectively.

Another benefit comes from multi-task learning is that compared to single task learning, it solves the tasks all at once with one model which significantly reduces computational cost and application complexity.

### 2.2. Dual-Loss Block

Traditionally, head pose is computed by estimating some key points from the target face and solving the 2D to 3D correspondence problem with a mean human head model. This is a fragile method because it relies entirely on landmark detection performance, the extraneous head model and an ad-hoc fitting step. Recently, there has been lots of progress [11] [12] [13] in using convolutional networks are focusing on regressing head pose - three Euler angles directly. However, it is hard to train neural networks to predict angles only using a single regression loss function. To mitigate this problem, we propose Dual-loss Block (DLB) to decompose the pose estimation task into pose classification and pose coarse-to-fine regression. By combining the two losses together, DLB takes both advantages of classification loss which can quickly converge into a small boundary and the regression loss which learns the residual to predict a more accurate value.

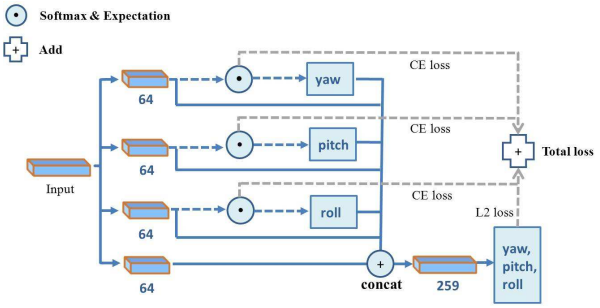


**Fig. 4**. Dual-Loss Block for Head Pose Estimation

As illustrated in Fig.4, given an input of a FC layer that contains high level features of the face, we apply four 1x1 Conv to aggregate and reorganize information into four paths. The first three paths are utilized for angle neighborhood classification of yaw, pitch and roll respectively. Once we get the classified result of each direction (each direction is divided into 12 groups in the range of $[-90^\circ, 90^\circ]$), we calculate the expectation of their softmax output to obtain a coarse angle prediction which will be further improved in the fourth path. The fourth path is designed for two goals. One is to aggregate global information and to look for relationship among yaw, pitch and roll. Another goal is to fit the residual mapping for the classification results and to eliminate the errors from classification loss. To do this, the aggregated and reorganized task-specific (yaw, pitch, roll) information and their normalized coarse angles are concatenated to the global features on the fourth path. As for loss, we employ cross entropy for classification and $L2$ for regression. Moreover, in order to reshape the loss function to down-weight easy angles (e.g., the one close to center) and thus focus training on harder ones [14], we extend the cross entropy loss by add a modulating factor $(1 - y_i)^\gamma$, where $y_i$ designates the models estimated probability for the class with label $y = 1$. The focusing parameter $\gamma$ smoothly adjusts the rate at which easy examples are down-weighted.

Finally, the head pose loss $\mathcal{L}_{hp}$ and all mentioned equations are as the followings:

$$y_i = \frac{e^{F(i)}}{\sum_{c' \in C} e^{F(c')}} \quad \forall c \in C \tag{3}$$

$$\mathcal{L}_{CE} = -(1 - y_i)^\gamma \log(y_i) \tag{4}$$

$$\mathcal{L}_{MSE} = \frac{1}{2p} \sum_{i}^{p} \left\| s - s' \right\|^2 \tag{5}$$

$$\mathcal{L}_{hp} = \mathcal{L}_{CE}^{yaw} + \mathcal{L}_{CE}^{pitch} + \mathcal{L}_{CE}^{roll} + \alpha \cdot \mathcal{L}_{MSE} \tag{6}$$

where $F(i)$ is the ith output of the last FC layer, $C$ denotes the set of angle neighborhood classification, $\mathcal{L}_{CE}$ is the cross entropy loss for single direction classification, $\mathcal{L}_{MSE}$ is the mean square error loss , and $p$ specifies three pose directions. Since gaze estimation is similar to head pose, we also adopt DBL in the gaze branch. The only difference is using two cross-entropy loss (2 dimensions) instead of three. Thus, the gaze loss is defined as: $\mathcal{L}_{gaze} = \mathcal{L}_{CE}^{x} + \mathcal{L}_{CE}^{y} + \beta \cdot \mathcal{L}_{MSE}$

### 2.3. Gaze Estimation

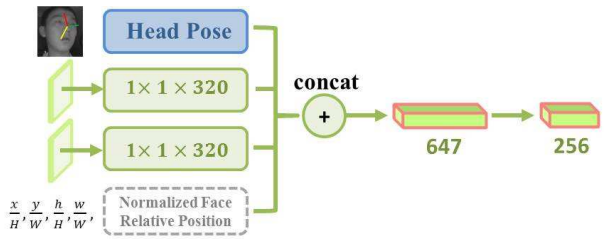


**Fig. 5**. Architecture of Objectcentric Aware Block

In a driver monitoring system, eye tracking is important to judge if the driver is paying attention to the road. Recently, some work [5] [15] [16] achieve better performance on person-independent gaze estimation, like [15] lowers the error to 1.71 cm on an iPhone and 2.53 cm on an iPad screen. However one drawback is that a fixed camera-screen relationship has to be satisfied, preventing these methods from being generalized to other devices. [5] introduced a way for unconstrained gaze estimation by normalizing the eye area. While conceptually elegant, it relies

entirely on accuracy of estimation of head pose and a generic 3D head model which introduces errors for any participant. Besides, separated operations also increases computational cost and system complexity. Our goal is to solve gaze estimation without making any assumption regarding the user or a 3D head model.

#### 2.3.1. *Objectcentric Aware Block*

We believe that satisfactory accuracy in gaze estimation can be achieved in constrained information like learned head pose. As Kundu mentioned in [17], object orientation can be egocentric (orientation w.r.t. camera), or allocentric (orientation w.r.t. object). Since orientation is predicted on top of a RoI feature-map (generated by cropping features on a box centered on the object), it is better to choose an objectcentric (allocentric) representation for learning.

To this end, we design an OAB (Objectcentric Aware Block) that enables the network to predict gaze angle with the help of the predicted head pose and the relative position of the face in the original image. To be more specific, as shown in Fig.5, after the high-level local information of the left and right eyes being collected by the network, OAB combines the information with learned head pose and the relative position of the driver face in the original image. Here, the relative position is used to indicate the location and size of the head within the frame, together with head pose increasing the network's objectcentric awareness. From the experimental result in Table.3, we can see that the OAB not only improves the accuracy of gaze estimation but also accelerates convergence speed during training. Since only 7(4+3) numbers are added to the input, the computational overhead increasing is negligible during inference.

## 3. EXPERIMENTS

**Table 1**. Face landmark evaluation on AFLW [18]

| Method | [0, 30°] | (30°, 60°] | (60°, 90°] | mean | std |
|---|---|---|---|---|---|
| SDM [19] | 3.75 | 5.55 | 9.34 | 6.55 | 2.45 |
| 3DDFA [20] | 5.00 | 5.06 | 6.74 | 5.60 | 0.99 |
| 3DDFA+SDM | 4.75 | 4.83 | 6.38 | 5.32 | 0.92 |
| HyperFace [21] | 3.93 | 4.14 | 4.71 | 4.26 | 0.41 |
| HF-ResNet | **2.71** | **2.88** | 3.19 | **2.93** | 0.25 |
| AIONet [10] | 2.84 | 2.94 | 3.09 | 2.96 | 0.13 |
| **DANet** | 2.93 | 2.94 | **3.08** | 2.98 | **0.08** |

**Table 2**. Head pose evaluation on AFLW2000 [22]

| **Method** | **Yaw** | **Pitch** | **Roll** | **MAE** |
|---|---|---|---|---|
| Dlib [23] (68 points) | 23.153 | 13.633 | 10.545 | 15.777 |
| 3DDFA [20] | 5.400 | 8.530 | 8.250 | 7.393 |
| FG [11] ($\alpha = 1$) | 6.920 | 6.637 | 5.674 | 6.410 |
| FG [11] ($\alpha = 2$) | 6.470 | 6.559 | 5.436 | 6.155 |
| **DANet** | **6.230** | **6.321** | **5.532** | **6.028** |

We perform experiments showing the overall performance of DANet on general face analysis datasets as well as DADrivers which is specific to driving environment. First, we evaluate our method on general face analysis datasets. As shown in Table.1&2, we conduct thorough comparisons between our model and existing state-of-the-art methods on AFLW [18] and its 3D version AFLW2000 [22], which are believed as classic and highly recognized datasets for general face analysis. We can see that our model achieves the best result on head pose estimation and a competitive result on face landmark detection. We argue that the reason why HF-ResNet (fourth line in Tab.1) is slightly better is because it adopts a heavyweight base network - ResNet which requires enormous computational cost.

**Table 3**. Ablation study of DBL and OAB on DADrivers.

| **Method** | **DBL Pose** | **DBL Gaze** | **OAB** | **Ldmk MAE** | **Pose MAE** | **Gaze MAE** | **CPU ms** |
|---|---|---|---|---|---|---|---|
| DANet | ✓ | ✓ | ✓ | **1.62** | **3.72** | **2.61** | ∼108 |
| DANet | ✓ | ✓ | | 1.62 | 3.72 | 3.22 | ∼106 |
| DANet | ✓ | | ✓ | 1.62 | 3.73 | 2.93 | ∼102 |
| DANet | | ✓ | ✓ | 1.82 | 3.98 | 3.01 | ∼101 |
| DANet | ✓ | | | **1.61** | 3.73 | 3.82 | ∼101 |
| DANet* | ✓ | | | **1.58** | **3.5** | - | ∼80 |
| DANet | | ✓ | | 1.83 | 4.02 | | ∼100 |
| DANet | | | | 1.82 | 4.03 | 3.83 | ∼95 |

**CPU:** Quad ARM Cortex-A57. **DANet*:** Without the gaze branch.

In Table.3, the ablation study clearly demonstrates the effectiveness of our proposed DLB (Dual-Loss Block) and OAB (Objectcentric Aware Block). For example, equipping with both two DLBs and OAB leads to best performance (first line in Table.3). Either getting rid of DBL Pose or DBL Gaze or OAB from the network results in a performance degradation (see line 2-4). However, it's worth noting that when we remove the gaze branch, the network DANet* outperforms the original network on the remaining face landmark and head pose branches. The current gaze branch has a negative effect on the other two branches, indicating that our network design needs further improvement.

**Fig. 6**. Visualization results of DANet on DADrivers in various scenarios (night, direct sunlight, wearing glasses and etc.)

## 4. CONCLUSION

In this paper, we propose a multi-task learning CNN framework DANet, which can simultaneously identify face landmark, head pose and gaze direction. With proposed DBL and OAB, the network not only achieves SOA results but also reduces the complexity of integration into automotive safety systems. It runs at around 10 fps on vehicle embedded systems which marks a momentous step for driver attention monitoring. More importantly, to facilitate other researchers, we publish our dataset FDUDrivers that might be the first comprehensive dataset and will serve as an invaluable resource in this field.

## 5. REFERENCES


[1] World Health Organization, "Global status report on road safety 2018," 2018.

[2] "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," *Traffic Safety Facts - Crash Stats*, 2015.

[3] A. Kumar and R. Patra, "Driver drowsiness monitoring system using visual behaviour and machine learning," in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, April 2018, pp. 339–344.

[4] Y. Qiao, Kai Zeng, Lina Xu, and Xiaoyu Yin, "A smartphone-based driver fatigue detection using fusion of multiple real-time facial features," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2016, pp. 230–235.

[5] X. Zhang, Y Sugano, M Fritz, and A Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans Pattern Anal Mach Intell*, vol. PP, no. 99, pp. 1–1, 2017.

[6] Mohamad-Hoseyn Sigari, Mahmood Fathy, and Mohsen Soryani, "A driver face monitoring system for fatigue and distraction detection," *International Journal of Vehicular Technology*, vol. 2013, 01 2013.

[7] Arthur Rumagit, Izzat Akbar, Mitaku Utsunomiya, Takamasa Morie, and Tomohiko Igasaki, "Gazing as actual parameter for drowsiness assessment in driving simulators," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, pp. 170–178, 01 2019.

[8] Mohammed Ghazal, Yasmine Abu Haeyeh, Abdelrahman Mohammad, and Sara Ghazal, "Embedded fatigue detection using convolutional neural networks with mobile integration," 08 2018, pp. 129–133.

[9] Serajeddin Ebrahimian, Ali Nahvi, Amirhossein Homayounfard, and Hamidreza Bakhoda, "Monitoring the variation in driver respiration rate from wakefulness to drowsiness: A non-intrusive method for drowsiness detection using thermal imaging," vol. 3, pp. 1–9, 01 2019.

[10] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa, "An all-in-one convolutional neural network for face analysis," *CoRR*, vol. abs/1611.00851, 2016.

[11] Nataniel Ruiz, Eunji Chong, and James M. Rehg, "Fine-grained head pose estimation without keypoints," 2018.

[12] Y. Yamaura, Y. Tsuboshita, and T. Onishi, "Head pose estimation for an omnidirectional camera using a convolutional neural network," in *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, June 2018, pp. 1–5.

[13] Massimiliano Patacchiola and Angelo Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognition*, vol. 71, 06 2017.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal loss for dense object detection," 10 2017, pp. 2999–3007.

[15] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M Bhandarkar, Wojciech Matusik, and Antonio Torralba, "Eye tracking for everyone," pp. 2176–2184, 2016.

[16] Erroll Wood, *Gaze Estimation with Graphics*, Ph.D. thesis, 10 2017.

[17] Abhijit Kundu, Yin Li, and James M. Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3559–3568, 2018.

[18] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 2144–2151.

[19] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*. jun 2013, IEEE.

[20] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 94–108, Springer International Publishing.

[21] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, Jan 2019.

[22] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 146–155.

[23] Roberto Valenti, Nicu Sebe, and Theo Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.