DPOSE: Online Keypoint-CAM Guided Inference for Driver Pose Estimation with GMM-based Balanced Sampling

Yuyu Guo Yancheng Bai Daiqi Shi Yang Cai Wei Bian

Alibaba Group Inc., China
{guoyuyu.gyy, yancheng.byc, daiqi.sdq, yangcai.cy, bianwei.ba}@alibaba-inc.com

Abstract

Human pose estimation (HPE) is an essential component of Driving Monitoring Systems (DMS) for real-time recognition of driving behavior. To achieve this, HPE is typically integrated with other tasks such as detection and head pose regression, into a single lightweight model that can be easily deployed on edge-side devices. However, oversimplified designs of lightweight HPE models may cause overfitting on generalized samples, rendering them unable to handle rare samples, particularly in the case of the dataset with the imbalanced distribution. In this paper, we propose an optimization scheme for a proprietary HPE task in DMS scenarios. Our method involves a pose-wise hard mining strategy to balance the pose distribution. Additionally, we introduce an online keypoint independent grad-cam loss, which constrains the gradient-based activation feature map of each keypoint prediction to its corresponding semantic region. We evaluate our approach using a benchmark dataset for DMS tasks and achieve outstanding results. Our code will be publicly available[1].

1. Introduction

Driver monitoring system (DMS) [8, 13] has become an indispensable component of modern automobile safety systems, owing to their ability to enhance driving safety and reduce the risk of accidents. In general, DMS relies on a combination of advanced sensors and processing algorithms to monitor driver behavior, ensuring that they remain attentive and fully in control of the vehicle at all times. A critical feature of DMS is human pose estimation (HPE), which enables the real-time capture of the driver's posture and movements [4, 17, 38], including the detection and tracking of their body joints and torso. This information is then utilized to analyze the driver's overall behavior and alert them if they exhibit any unsafe driving behaviors that could pose a risk to themselves or others on the road [9, 17].

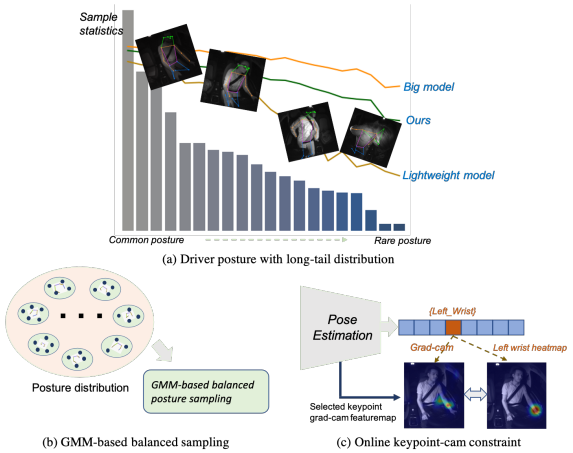

Figure 1. The illustration of driving posture in DMS: (a) the common posture in driving scenario is characterized by an unbalanced distribution. (b) our proposed pose-wise balanced sampling strategy via GMM cluster. (c) online keypoint-cam constraint to guide the model focuses on specific keypoint regions.

Human pose estimation (HPE) [2, 19, 34] is a fundamental task for human-centered relevant studies, *i.e.*, human behavior study [3, 23], action interaction [39, 46] and anomaly event recognition [18, 22, 31], *etc.* In the past decades, the great success of deep learning techniques has greatly boosted the performance of human pose estimation [7]. However, most of the existing HPE approaches highly rely on complex network design and large model parameters to improve their performance. For instance, the state-of-the-art HPE model, HigherHRNet [6], requires over 150GMACs computational operation per frame, which severely impedes its implementation on edge devices. To address this challenge and promote the practical application of HPE models, the research on lightweight models for HPE has gradually attracted attention. As per the literature,

[1] https://github.com/yyguo0536/DPOSE.git

Wang et al. [43] design an efficient architecture including fusion deconvolution head and large kernel convolutions for pose estimation. Maji et al. [21] develop YOLO-Pose framework which integrates human detection and pose estimation together to save inference time. Additionally, some researchers have explored the loss functions to improve the accuracy of light HPE, *i.e.,* RLE loss [14], OKS loss [21].

Although the above-mentioned methodologies have improved the performance of lightweight models at the wild HPE task, they are not well-suited for HPE in driving scenarios. For driving pose estimation, the long-tail posture distribution is a critical challenge that can greatly affect the performance of body pose estimation, particularly for lightweight models. This is because the lightweight model is suffered from limited data modeling capability which may cause the model prone to overfitting to common samples, as shown in Fig. 1 (a). For instance, when the posture distribution is uneven, such as the long-tail distribution in Fig. 1 (a), the light model may achieve the accurate estimation in the common samples with high-frequency happening, but performs poorly in the rare pose samples. Therefore, it is essential to address the imbalanced distribution problem to improve the generalization capabilities of light models in pose estimation. In addition, regression-based one-stage HPE methods also suffered from the problem of extremely sharing features between different keypoints caused by the short-range specific keypoint head design. To tackle this issue, some feature decoupling approaches are presented [11, 41]. However, most of them decouple the features by adding extra layers or extending the specific head layer, and these designs are arduous to employ on the edge device. In this work, we present an online keypoint-cam constraint to disentangle the strong shared features between the keypoints. The design regards the prediction of each keypoint as an independent target detection and classification task [42], therefore the online Grad-CAM constraint [15] can be used to guide the model pay attention to the area corresponding to each key point.

In summary, we present a simple yet effective HPE framework which specific optimized for DMS scenarios, termed DPOSE. Our contributions can be summarized as follows:

1. We present a uniform sampling strategy for the uneven distribution of driving poses. The GMM model is leveraged to perform clustering statistics on the driving posture, and the observed coefficients for each cluster are used for uniform sampling during HPE training.

2. We introduce an online keypoint-cam guidance for each keypoint inference. In our method, each keypoint from different body parts is regarded as a multi-object detection and classification task. Then the gradient-based classification activation map is leveraged to decouple the casually shared features in keypoint-wise.

3. We evaluate the proposed framework on the SOTA benchmark dataset, and achieve outstanding performance.

2. Related Works

Although human pose estimation is crucial to DMS, existing relevant works routinely utilize the proposed HPE models directly [9, 17], instead of optimizing them for in-vehicle scenarios. In the following, we partitioned the related works section into three categories that we deemed relevant to this research: (1) The top-down based HPE approaches; (2) bottom-up based HPE approaches, and (3) CAM applications for computer vision.

2.1. Top-down HPE

The top-down HPE commonly employs a hierarchical scheme to detect humans first, and then estimate a single human pose sequentially from the image. The popular detection methods, i.e., YOLO [28], and Faster-RCNN [29], are coupled with top-down HPE. This allows the model to concentrate on improving the accuracy of keypoint locating. There are various works about top-down HPE presented in the past decades, that significantly boost the performance of HPE. For instance, Hourglass [25], PoseNet [27], SimpleBaseline [44], and HRNet [35], etc., have demonstrated their superior ability for HPE. However, this two-stage paradigm is computationally intensive and significantly increases the computation time. This causes them difficult to implement practically in edge devices.

2.2. Bottom-up HPE

The bottom-up HPE achieved rapid development in the past decades. Typically, classical bottom-up based methods capture the potential identity-free keypoints first and then assign them to individual humans by the relative relationship between the keypoints. For the recent optimization of bottom-up HPE, the researchers mainly focus on the assignment of the keypoints to individual person and solving the scale variance of different individuals in the same scene. For instance, Cao et al [5] pioneeringly presented DeepPose that associates the extracted keypoints via predicting part affinity fields. Kocabas [12] then introduced PRN network, which combines the detection task and pose estimation task together to replace the keypoint assignment process. Luo et al [20] presented a scale-adaptive heatmap regression (SAHR) method to normalize the variant instance on large scale.

One-stage HPE: To further boost the computational efficiency of bottom-up HPE, the researchers propose one-stage solutions. For instance, CenterNet [50] and DirectPose [37] are the pioneering work for single-stage HPE that

directly regress the object coordinate position end-to-end. On this basis, researchers have proposed several methods to improve the one-stage regression HPE performance, especially for lightweight models. Maji et al. [21] modify the yolo-series network to YOLO-Pose by adding the extra keypoints head and designing OKS loss for keypoint regression.

Feature decoupling for HPE: Indeed, the HPE task can be inherently regarded as a multi-object detection task in that each joint is an individual instance. Thus, each keypoint is expected to have its specific feature representation, rather than using universal shared features between keypoints. Current researchers improve the feature expression ability of each keypoint via feature decoupling strategies. For instance, Wei et al. [36] statistically analyzed the correlation between different body parts from the training dataset using mutual information and accordingly split them into several groups with separate prediction heads. Geng et al [11] presented an adaptive convolution design to disentangle the keypoint representations respectively. KAPAO [24] regards the specific keypoints and set of semantically relevant keypoints (i.e., pose) as objects within a single-stage detection framework.

2.3. CAM Applications

CAM [49] is able to generate the coarse class activation maps highlighting the visual attention region that is responsible for the network's decision. It is thus widely used in the field of model interpretability [47] and weakly supervised learning [1]. Grad-CAM [33] is the extension of CAM which consider the gradient information of the specific classification. Many existing methods employ the Grad-Cam in the field of weakly supervised learning because it can roughly locate image foreground regions only relying on classification information. For instance, Li et al [15,16] presented a novelty trainable guided attention inference framework to segment the target object only supervised classification labels to generate the corresponding network attention map. Wang et al [40] introduced a CAM-loss to improve the discriminative feature representation of the backbone. Xie et al. [45] recently employ the class-agnostic activation map to contrastive learning, to further reduce the requirements of the supervision information of detection/segmentation tasks. Inspired by Grad-CAM's success on computer vision tasks, we propose a specific keypoint-cam guided decoupling constraint similar to [16], to improve the representative capability of the backbone network.

3. Method

Fig 2 illustrates the overall framework of our proposed method for light HPE model training. In the following section, we first introduce the architecture of our network in Sec. 3.1. Then, a GMM-based pose statistical method is presented in Sec. 3.2 to reduce the distribution bias of the training dataset in posture-wise. Thirdly, we give the details of our backbone feature decoupling design in Sec. 3.3. Finally, in Sec. 3.4 we summarize our training loss of the whole framework.

3.1. Architecture

Let I denote the input target image that includes a driver with the specific posture in the driving scenario, and K denotes the number of interest keypoints. The learnable parameters of the backbone network can be denoted as θ. The backbone θ extracts the feature maps $F^i_{c\times h\times w}$ in different scales from I, where i indicates the $i-th$ layer of backbone and $c\times h\times w$ denote the channel number and spatial resolution respectively. Following the backbone network, the specific head layer ϕ_d, ϕ_p is to achieve the detection task and pose estimation task separately.

To improve the computational efficiency possibly, we consolidate the human detection and pose estimation task in a single-stage to eliminate the reliance on the keypoint association module which is similar to the prior [21]. Image I firstly feed into the backbone network θ to extract the fundamental shared features. Subsequently, two individual head layers are connected to the output of the backbone. In the training stage, the positive potential samples are selected via a strategy like YOLOX which is anchor-free framework. The formulation of the detection and pose estimation can be intuitively denoted as:

$$\left\{\mathcal{B},\mathcal{P}\right\} = \mathcal{D}(I|\phi,\theta) \tag{1}$$

where $\mathcal{B}$ and $\mathcal{P}$ stand for the driver bounding box and pose output, respectively. The bounding box output $\mathcal{B}$ is composed by (x_{center}, y_{center}, x_{width}, y_{height}, c_b), and pose coordinate position $\mathcal{P}$ is represented by $\left\{(x_k, y_k, c_k), k=1,...,K\right\}$, where K indicates the number of keypoints. Similar to the prior bottom-up methods, we also adopt the center offset-based method to estimate the bounding box and pose accurate position.

In our framework, a GMM-based posture statistical distribution model is first trained and utilized to balance sampling the training dataset, which can mitigate the long-tail data distribution problem and improve the model performance on rare driving postures. Secondly, to further decouple the shared features between each keypoints, the class-agnostic activation maps can be leveraged to constrain and reduce the incorrect causally associated relationship with each other keypoint. Ideally, the corresponding gradient-guided CAM is supposed to activate the foreground image region while suppressing background regions. Thus, a pixel-wise posture online grad-cam is employed for backbone feature extraction, to disentangle the selected foreground and relative background features. As shown in Fig 2, the specific keypoint prediction is randomly selected and

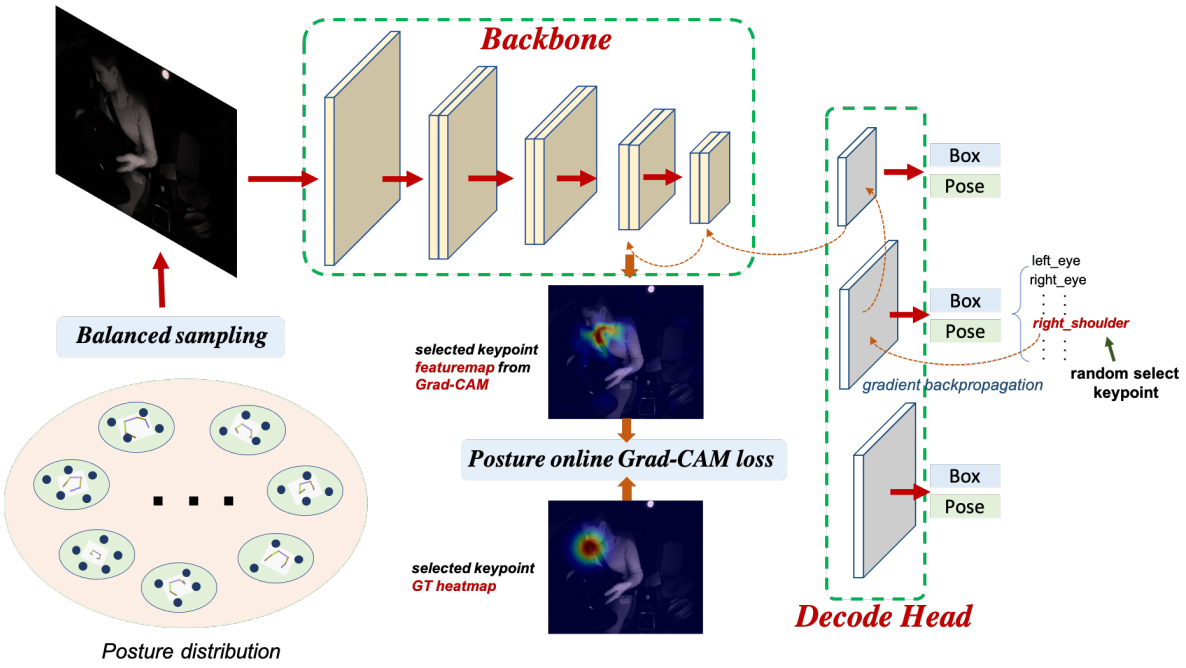

Figure 2. Overall framework of the proposed DPOSE.

computed its corresponding loss, then the gradient map G^i in the target interest layer G^i via loss backpropagation wise-product to the feature map F^i to generate the corresponding CAM map.

3.2. GMM-based Clustering for Pose Balance Sampling

As shown in Fig. 1, the driver's posture is not uniformly distributed in the daily driving scenarios, which can significantly limit the performance of human pose estimation models, especially for light models. This is because the light model is suffering from limited data modeling ability, which may overfit into the common samples, i.e., general posture in driving HPE tasks. Thus, it is important to balance the posture distribution in the training dataset while including the diverse range of postures to improve the HPE performance. To this end, we present a pose-wise balanced sampling strategy to mitigate the negative impact on the performance of HPE. Considering that there is a relative similarity or even repetition between human postures, the driver's posture can be summarized into several subgroups. Such posture groups can be represented by a mixture of Gaussian models. In this way, various types of postures can be parameterized and statistically counted, then the uniform sampling strategy can be performed according to the statistical number during the model training stage. Furthermore, it also benefits for hard sample mining to help improve the model performance.

In order to effectively cluster the body poses and reduce the interference caused by the position change of the human body in the image, the coordinate position of each key point is first normalized into an offset relative to the center of the human body frame as follows.

$$\hat{\mathcal{P}}_k = \mathcal{P}_k(x_k, y_k) - \mathcal{B}(x_{center}, y_{center}) \tag{2}$$

The normalized pose representation can be regarded as a random vector following a multivariate Gaussian distribution with mean μ and a covariance matrix Σ. Let $\omega_j > 0$ indicates the dominance of an observation $\mathcal{P}^j$ as suggested in [10]. The GMM model is then trained with normalized postures and clusters the posture samples into N categories as:

$$\{\mathcal{P}_j\} = \mathcal{G}\{\mathcal{P}|\mu_j, \Sigma_j, \pi_j\}, j = 1, ..., N \tag{3}$$

where μ_j, Σ_j, π_j are the mixtures parameters: μ_j and Σ_j are the feature distribution representation of $j-th$ component, π_j is the mixing coefficients satisfying $\sum_{j=1}^{N} \pi_j = 1$. The observed data is sampled via the weights provided by the GMM model.

3.3. Keypoint-CAM Decoupling Constrain

As we discussed in Section 2, the feature decoupling between the different joint keypoints can improve the HPE accuracy. However, current feature decoupling designs require to increase the network layer to split the keypoints' features in physics which increases the complexity of the network. Besides, these methods do not improve the feature

expression of the backbone. To solve this issue, we design a separate pose Grad-CAM constraint directly working on the backbone feature maps in a regularized bootstrapping manner. As detailed description in Fig. 2, each keypoint estimation can be regarded as a separate regression task, thus grad-cam mechanism can be utilized to extract their individual attention region on the shared feature maps from the backbone. In this way, the different keypoint is able to localize their own interest region to disentangle the feature maps.

In this paper, we adopt the online Grad-CAM mechanism similar to [15] to generate the trainable attention map according to the input samples within each training iteration. Each keypoint's gradient is computed from its corresponding visible ground-truth mask c_k, and we can subjectively capture the gradient map from the interest layer of the network. The computed gradient map can be regarded as the neuron importance weights responsible for the network output of the keypoint. The computation of the gradient map is defined as:

$$\mathcal{G}_i^{c,u,v} = \frac{\partial(\hat{c}_k - c_k)}{\partial f_i^{c,u,v}} \tag{4}$$

where $\mathcal{G}_i^{c,u,v}$ denotes the gradient map in $i-th$ layer, and $\{c, u, v\}$ indicate the channel and spatial index which is consistent to the interest feature maps $f_i^{c,u,v}$. Once obtained the importance weights represented by $\mathcal{G}_i^{c,u,v}$, it is then element-wise product to the feature maps to generate the attention map covered on the keypoint-affected region.

$$\mathcal{A}_i = ReLU\Big(\sum_{c=1}^{C} \mathcal{G}_i^c \odot f_i^c\Big) \tag{5}$$

We should note that the backward of keypoint confidence loss $(\hat{c}_k - c_k)$ will not update the network parameters, and it is only leveraged to obtain the importance weights to generate the attention map. We then normalize the attention map to range $(0, 1)$ using sigmoid function:

$$\mathcal{A}_i = \frac{1}{1 + exp\big(-\omega(\mathcal{A}_i - \sigma)\big)} \tag{6}$$

where ω and σ denote the thermal weights to adjust the scale and contrast of the attention map. To regularize bootstrap the attention map focuses on the specific body part region, we then generate the corresponding keypoint heatmap from its position labels and employ it as a constraint to the attention map. The generated keypoint heatmap $\mathcal{H}_k$ is defined as:

$$\mathcal{H}_k = exp\Big(-\frac{\Big((x_k, y_k) - \mu_k(x, y)\Big)^2}{2\sigma^2}\Big) \tag{7}$$

In order to enforce the network concentrate on the corresponding body part of $k-th$ keypoint, the termed attention loss $\mathcal{L}_{att}$ is denoted as:

$$\mathcal{L}_{att} = \|\mathcal{H}_k - \mathcal{A}_k\|_2 \tag{8}$$

By minimizing the attention loss $\mathcal{L}_{att}$, the network learns to focus on the different triggered visual regions contributing to the keypoints without any extra network parameters.

3.4. Training Loss

In this section, the details of the loss function are introduced to train the model. Overall, the proposed network consists of two tasks: human detection and pose estimation. Note that, the detection task is mainly leveraged to filter the potential positive sample for HPE. And we clarify the loss design in two individual parts.

Detection loss For the detection task, the bounding box coordinate is regressed in the manner of (x^m_{center}, y^m_{center}, x^m_{width}, y^m_{height}), which is the position offset relative to the center and the height and width of the target bounding box. The ambition of detection is to maximize the Intersection over Union (IoU) between the predictions and the ground-truth labels. Therefore, IoU based loss function including its variant form is popular and widely used for detector training, i.e., GIoU [30], DIoU [48]. In our implementation, we utilize the DIoU loss to supervise the human detection training. Its formulation is denoted as:

$$\mathcal{L}_{box} = (1 - DIoU(\mathcal{B}, \hat{\mathcal{B}})) \tag{9}$$

In addition to the bounding box regression, the classification-based confidence prediction is supervised as:

$$\mathcal{L}_{b_c} = \sum_{k=1}^{K} BCE(\hat{c}_k, 0/1) \tag{10}$$

Pose loss Recently, there are several excellent loss functions about pose keypoints proposed, i.e., RLE loss, and OKS loss. Indeed, most of these approaches improve the performance of the HPE model by modeling the distribution of each keypoint. For instance, RLE loss utilized the flow model method to estimate the exact uncertainty of the keypoint with its specific mathematical distribution. OKS loss assumes that each keypiont's possible distribution conforms to Gaussian distribution, and adopts the uncertainty of the keypoint derived from prior knowledge to constrain the HPE learning. However, this distribution-based constraint can not effectively reflect the distance error when the predicted coordinate bias is beyond the prior range. Considering the limited modeling capability of the lightweight model, we leverage the OKS loss for pose coordinate training. Its formulation is derived as:

$$\mathcal{L}_{kpts} = \sum_{k=1}^{K} \left(1 - exp\left(\frac{(\hat{x}_k - x_k)^2 + (\hat{y}_k - y_k)^2}{2s^2\sigma_k} \right) \right) \tag{11}$$

where $(\hat{x}, \hat{y})$ is the keypoint coordinate prediction and (x, y) is ground-truth label. σ_k denotes the prior uncertainty value which is the same as the evaluation metric of the previous works. In addition to the coordinate output, a confidence estimation is accompanied by presenting the target object have the corresponding keypoint or not as Eq 12.

$$\mathcal{L}_{p_c} = \sum_{k=1}^{K} BCE(\hat{c}_k, 0/1) \tag{12}$$

In summary, the total loss includes the above presented loss in this section and our proposed decoupling regularized constrain $\mathcal{L}_{att}$. The final training loss is:

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_{att} + \lambda_2 \mathcal{L}_{box} + \lambda_3 \mathcal{L}_{b_c} + \lambda_4 \mathcal{L}_{kpts} + \lambda_5 \mathcal{L}_{p_c} \tag{13}$$

where $\lambda_1 = 5, \lambda_2 = 1, \lambda_3 = 1, \lambda_4 = 1, \lambda_5 = 1$ in our experimental setting.

4. Experiments

4.1. Materials and Evaluation Metrics

We demonstrate our approach with the public benchmark datasets for the driving scene: *Drive&Act* dataset[2] which is a state-of-the-art multi-modal benchmark for action recognition in automated vehicles. The dataset consists of 12 hours of video data in 29 long sequences recorded from 17 different persons. We randomly selected 33.6K images including the driver bounding box and 17 keypoints in coco format as the training dataset and test dataset. Among them, the training dataset contains around 26.9K images from 13 persons, while the test dataset contains the left 6.7K images from 4 persons. We follow the standard evaluation metric and use OKS-based metrics for MPPE. We report average precision with different thresholds: mAP, AP0.75, and mAR.

4.2. Implementation Details

We compare our proposed method with several different categories of state-of-the-art pose etimation methods, including heatmap-based methods, i.e., light-openpose [26], movenet [32], regression-based method, Yolo-Pose [21]. We implement all the compared methods following their default parameter settings in the same environment, including the same training set, and the same testing set. We adopt Yolo-Pose as our baseline model cause of its high computational efficiency.

The training data is augmented by random image mirroring, color contrasts random transformation, etc. following the previous similar research [21]. We implemented the networks using the Pytorch library and trained on Nvidia 1080Ti hardware. We utilized the Adam optimizer to train the model 350 epochs and the initial learning rate is set to $1 \times e^{-4}$ and dropped by a factor of 10 after 250 epochs.

4.3. Ablation Study

We conduct an ablation study for the critical component in our proposed framework. To effectively discuss the impact of the proposed GMM-based balanced sampling strategy and the different combinations of loss functions on the accuracy of pose estimation, we individually analyze the presented method with different combination schemes as shown in Table 1. Firstly, we only implement the baseline model with detection and pose loss, and then we gradually increase our proposed GMM-based sampling strategy and pose cam loss to compare the performances. Note that, the cluster number setting is 20 in the following GMM based balanced sampling in Table 1.

Table 1. The quantitative performance of driving pose estimation on the *Drive&Act* dataset for different ablation settings.

Baseline	Balanced sampling	Keypoint-CAM	mAP	AP0.75	mAR
✓	-	-	0.761	0.855	0.847
✓	✓	-	0.824	0.930	0.881
✓	-	✓	0.798	0.897	0.869
✓	✓	✓	**0.836**	**0.949**	**0.893**

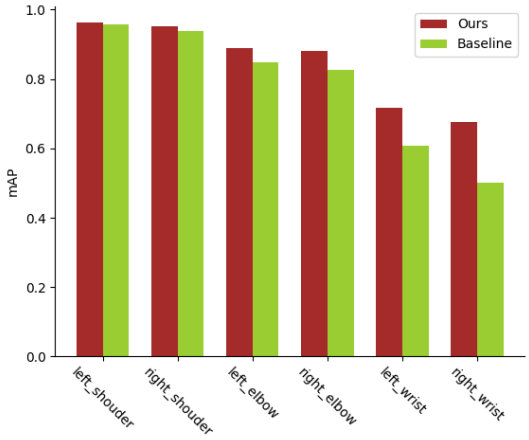

Figure 3. The comparisons result on specific keypoints.

Compared to the baseline model, the proposed balanced sampling strategy greatly improves the accuracy of the HPE

[2]https://driveandact.com/

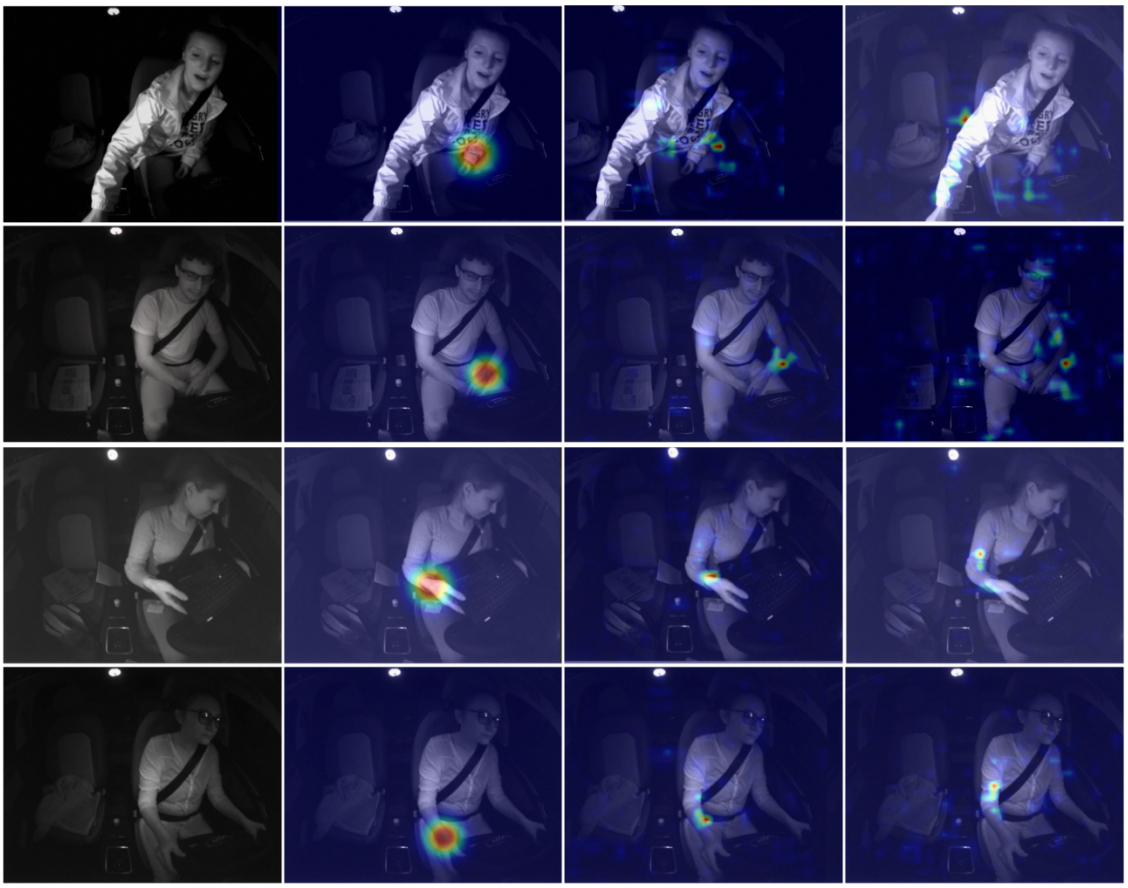

Figure 4. Qualitative results of the selected samples from *Drive&Act*. It visually compares the Grad-CAM feature map from specific keypoints with the baseline model and our proposed method.

task, which increases the mAP to around 0.75. Meanwhile, only leveraging online keypoint-cam constraints increase the mAP metrics around 0.37 benefitting from its specific keypoint attention guidance and feature decoupling ability for the shared features from the backbone. Furthermore, the simultaneous use of balanced sampling strategy and online keypoint-cam guidance can further improve the accuracy of HPE.

In practical applications, human limbs have a wider spatial variation, which is relatively more difficult to predict. Fig. 3 illustrates that our proposed method is more effective for boosting the keypoints with large variance than the baseline model.

4.3.1 GMM-Based Balance Sampling

Regarding the above results that the balanced sampling strategy significantly impacts the performance of HPE, we then discuss the influence on different settings of balanced sampling as presented in Table 2. The number of categories in GMM clustering for the driving posture is artificially set. In order to verify the best set of cluster num, we increase the number setting from 8 to 20 with 4 steps. Initially, the quantitative performance of HPE is consistent with the cluster number increasing. However, the HPE accuracy decreases with the cluster number increase after 16. We attribute this to that excessive pose sampling strategy may affect the performance of the model on common distribution samples.

Table 2. The quantitative performance of driving pose estimation on the *Drive&Act* dataset for different GMM cluster settings.

	mAP	AP0.75	mAR
Baseline	0.761	0.855	0.847
Cluster:8	0.812	0.914	0.879
Cluster:12	0.837	0.922	0.885
Cluster:16	**0.851**	**0.935**	**0.898**
Cluster:20	0.824	0.930	0.881

Table 3. Driver pose estimation results on the *Drive&Act* dataset for the comparison methods.

	Method	Backbone	Input Size	Params(M)	Gflops(G)	mAP	AP50	AP75	mAR
Heatmap-based	MoveNet	MobileNetV2	256×256	1.91	**0.86**	0.874	0.975	0.938	0.926
	LightOpenPose	MobileNetV1	368×460	2.72	14.62	0.887	0.925	**0.980**	0.906
Regression-based	Yolo-Pose	DarkNet-0.125	480×480	0.63	1.26	0.795	0.978	0.881	0.866
	Yolo-Pose	DarkNet-0.125	640×640	0.63	2.23	0.761	0.978	0.855	0.847
	Yolo-Pose	DarkNet-0.25	640×640	2.47	8.24	0.888	0.988	0.952	0.921
	Ours	DarkNet-0.125	480×480	**0.63**	1.26	0.871	0.978	0.963	0.917
	Ours	DarkNet-0.125	640×640	0.63	2.23	0.851	0.987	0.935	0.898
	Ours	DarkNet-0.25	640×640	2.47	8.24	**0.902**	**0.989**	0.965	**0.939**

4.3.2 Keypoint-CAM Visulization

In this section, we visually compare the feature maps between the cam-constrained and baseline models as shown in Fig. 4. The left two columns indicate the input image and selected keypoints heatmap separately, and the right two columns present the Grad-CAM-based feature map visualization of our method and baseline model. In Fig. 4, the top two rows show the left wrist keypoints and the bottom two rows show the right wrist keypoints, which we consider the wrist estimation is relatively important for the DMS task. With the keypoint-cam constraint, the HPE model concentrates on the specific-affected regions related to the target keypoint, and filters out the irrelevant information. This enhances the anti-interference ability of the model, thereby improving the keypoint accuracy of the HPE model.

4.4. Comparison with SOTA Methods

We have compared our proposed method to the existing heatmap-based and regression-based HPE approaches. As shown in Table 3, our method achieved competitive results compared with the existing heatmap-based approaches, while our method leverages fewer model parameters. We should note that the MoveNet in our reimplementation requires the prior body bounding box to preprocess the input images. In addition, our method outperformed the compared SOTA regression-based methods. Overall, the proposed method retains the lightweight design of the model while ensuring HPE accuracy, especially for DMS-related tasks, which can be combined with other related tasks into a unified model without affecting other task indicators.

5. Discussion and Conclusion

We present a simple yet effective framework for driving posture estimation. Our main findings are that our method: 1) effectively captures the diverse driving posture in the DMS task via the distribution statistical sampling strategy, which reduces the negative effect of the long-tail distribution for HPE; 2) is robust and stable to capture the driver's pose and not susceptible to irrelevant information benefitting from the individual online keypoint-cam guidance. 3) is readily integrated with other downstream tasks in DMS, i.e., detection, head pose regression, etc., and not barely increases the model computation.

In our framework, the balanced sampling scheme greatly augments the hard and rare driving posture and improves the model performance. As shown in Fig. 1, we have found that the accuracy of the HPE prediction is strongly correlated with the frequency of the posture in the training dataset. This is attributed to the regression-based HPE approaches highly relying on the various diversity of the posture, otherwise, the model is prone to overfitting to some common poses. More specifically, the coordinate regression model does not establish a mapping relationship between the image domains, but a high-dimensional nonlinear mapping relationship between the image domain and the coordinate domain, which makes the lightweight model more susceptible to the influence of uneven data distribution and suffering from the common-mean posture overfitting problem cause its limited modeling ability. Our balanced sampling design can effectively reduce the above issues and ensures that the pose estimation is more stable and robust.

In the visual comparison of feature maps, the keypoints' feature are readily coupled together, which may lead to the model learn the wrong association information between each other. This also is a critical reason why the model is prone to overfit to the common group of postures. In contrast, the keypoint-cam constraint decouples the feature association by guiding the model concentrates on the potential attention region of each keypoint, as shown in Fig. 4.

In conclusion, we introduce optimized sampling for the problem of uneven distribution of attitude samples, and put forward targeted constraints on the overfitting problem that is prone to occur in lightweight models. In addition, our solution can be quickly and conveniently integrated with DMS-related downstream tasks with rarely additional computation to the model. We suggest our proposed balanced sampling strategy and online-cam constraint have a general contribution to the regression task of the lightweight model.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 3

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 1

[4] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4661–4670, 2017. 1

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 1

[7] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 1

[8] Yanchao Dong, Zhencheng Hu, Keiichi Uchimura, and Nobuki Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE transactions on intelligent transportation systems*, 12(2):596–614, 2010. 1

[9] Mohit Dua, Ritu Singla, Saumya Raj, and Arti Jangra. Deep cnn models-based ensemble approach to driver drowsiness detection. *Neural Computing and Applications*, 33:3155–3168, 2021. 1, 2

[10] Israel Dejene Gebru, Xavier Alameda-Pineda, Florence Forbes, and Radu Horaud. Em algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2402–2415, 2016. 4

[11] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021. 2, 3

[12] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018. 2

[13] Arief Koesdwiady, Ridha Soua, Fakhreddine Karray, and Mohamed S Kamel. Recent trends in driver safety monitoring systems: State of the art and challenges. *IEEE transactions on vehicular technology*, 66(6):4550–4563, 2016. 1

[14] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11025–11034, 2021. 2

[15] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 2, 3, 5

[16] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):2996–3010, 2019. 3

[17] Taiguo Li, Tiance Zhang, Yingzhi Zhang, and Liben Yang. Driver fatigue detection method based on human pose information entropy. *Journal of advanced transportation*, 2022, 2022. 1, 2

[18] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 1

[19] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1159–1168, 2018. 1

[20] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264–13273, 2021. 2

[21] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 2, 3, 6

[22] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 1

[23] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 1

[24] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. *arXiv preprint arXiv:2111.08557*, 2021. 3

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 2

[26] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv preprint arXiv:1811.12004*, 2018. 6

[27] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017. 2

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[31] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020. 1

[32] Votel Ronny and Li Na. Movenet. https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html. 6

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[34] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021. 1

[35] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[36] Wei Tang and Ying Wu. Does learning specific features for related parts help human pose estimation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1107–1116, 2019. 3

[37] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*, 2019. 2

[38] Helena R Torres, Bruno Oliveira, Jaime Fonseca, Sandro Queirós, João Borges, Nélson Rodrigues, Victor Coelho, Johannes Pallauf, José Brito, and José Mendes. Real-time human body pose estimation for in-car depth images. In *Technological Innovation for Industry and Service Systems: 10th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2019, Costa de Caparica, Portugal, May 8–10, 2019, Proceedings 10*, pages 169–182. Springer, 2019. 1

[39] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 1

[40] Chaofei Wang, Jiayu Xiao, Yizeng Han, Qisen Yang, Shiji Song, and Gao Huang. Towards learning spatially discriminative feature representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1326–1335, 2021. 3

[41] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 2

[42] Lan Wang and Vishnu Naresh Boddeti. Do learned representations respect causal relationships? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–274, 2022. 2

[43] Yihan Wang, Muyang Li, Han Cai, Wei-Ming Chen, and Song Han. Lite pose: Efficient architecture design for 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2022. 2

[44] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2

[45] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022. 3

[46] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005, 2019. 1

[47] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018. 3

[48] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020. 5

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2