# MSP-DISK: Naturalistic and Diverse In-Vehicle Database for Joint Pose and Seat Belt Detection

Isaac Brooks[1], Susmitha Gogineni [1], Sumit Jha[1], Soumitry Jagadev Ray [2], Rajesh Narasimha [2], Naofal Al-Dhahir[1], and Carlos Busso[1]

***Abstract*— Driver monitoring systems improve the safety of the car by alerting the driver when unsafe behavior is detected. To understand the driver and passengers, it is important to estimate the location of their body keypoints and detect the presence of seat belts. This information can be used to ensure that the driver is facing forward and all passengers are properly wearing their seat belts. Modern computer vision methods perform well in the task of human pose estimation for varied environments. However, in-vehicle scenarios present unique difficulties due to varying lighting conditions, camera placement, and lack of specific in-domain training data. We propose the *diverse in-vehicle seat belt and driver keypoint* (MSP-DISK) dataset. This corpus consists of images sourced from online video-sharing websites that contain a variety of vehicle interiors, subjects, illumination conditions, camera angles, and camera qualities. Each image contains annotations for the visible upper-body keypoints for the driver and all passengers in the vehicle and a binary mask indicating the location of the seat belt for all subjects. We implement a lightweight computer vision model jointly trained on this dataset to detect seat belts and body keypoint locations. For seat belt detection, we obtain an *intersection over union* (IoU) equal to 27.8%. For the body keypoint detection, we obtain an F1-score of 0.773.**

## I. INTRODUCTION

Unsafe driving behaviors such as distracted driving and improper seat belt usage lead to thousands of deaths and injuries every year. In 2020, 13% of all vehicle accidents in the U.S. involved distracted driving, causing over 3,000 deaths and 300,000 injuries according to the the *National Highway Traffic Safety Administration* (NHTSA) [28]. Likewise, proper usage of seat belts saved over 14,000 lives in the U.S. in 2017 [27]. Failing to comply with safe driving guidelines can result in avoidable fatalities. The implementation of a *driver monitoring system* (DMS) that detects unsafe driving behaviors [21], [22] and alerts the driver when unsafe behavior is detected [30] can be instrumental in preventing such deaths.

Existing DMS components include seat belt latch detection and warning mechanisms, which are mandated in all new passenger vehicles in the U.S. [7]. However, these systems can be tricked by latching the seat belt and wearing the belt behind the shoulder. Also, there is no existing standard DMS component that measures driver distraction. A solution is to use artificial intelligence to detect unsafe driver behavior. An artificial intelligence-based system can be robust to variations in driver behavior to more accurately detect unsafe situations. Such a system may combine the data collected from various sensors within the vehicle to predict whether the driver is paying adequate attention to the road and whether all seat belts are properly fastened. A dashboard-mounted camera facing the vehicle cabin, for example, can monitor the driver and passengers and capture visual information relating to pose and seat belt status.

Recent advances in 2D *human pose estimation* (HPE) have led to models achieving state-of-the-art performance (see for example results on the MSCOCO Keypoints Challenge [2], [3], [24]). However, conventional HPE approaches focus on a broader domain than in-vehicle scenarios. For an intelligent DMS, we want an HPE model that can accurately detect the pose of human subjects in the vehicle to determine whether the driver is paying attention to the road. For example, the rotation of the driver's shoulders may indicate that the driver is facing the back of the vehicle instead of the road. The detection of the hands can also indicate whether the drivers have their hands on the steering wheel [20]. Using deep learning techniques, a model trained on human pose data can learn to accurately predict the pose of human subjects within an image. Training a model on domain-specific data from an in-vehicle camera can prepare the model for the challenges of the in-vehicle domain, such as illumination variations (e.g., glare, low-light conditions, and abrupt illumination changes as the vehicle moves), non-ideal camera angles for the task, and occlusions within the vehicle.

Unfortunately, while several studies have collected pose and/or seat belt data in an in-vehicle domain, few datasets are publicly available. Furthermore, the available resources are often videos recorded in the same vehicle, using the same camera and experimental setup. For the safety of conducting these in-person experiments, it is necessary to control as many aspects of the experiment. Therefore, many studies that collect visual in-vehicle data perform controlled trials where each subject drives the same vehicle while being recorded. One consequence of using the same vehicle is that the vehicle interior, seat belt color, and seat belt texture are consistent for all the collected images. A model trained on this data may overfit to the representation of a seat belt in a particular vehicle but perform worse on seat belt detection in another vehicle. To remedy these issues, we introduce the *diverse in-vehicle seat belt and driver keypoint* (MSP-DISK) dataset, a publicly available dataset of diverse, in-vehicle images from

This work was supported by NSF under grant IIP-1950249.

[1]Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080 {isaac.brooks, sxg155930, sumit.jha, aldhahir, busso}@utdallas.edu

[2]EdgeTensor, Dallas, TX 75230 {soumitryjray, rajeshnarasimha}@edgetensor.com

a variety of video sources.

The MSP-DISK dataset includes 7,704 images from 542 different videos containing upper-body keypoint locations for the driver and all passengers in the vehicle. The upper-body joints are the most often visible joints when using a dashboard-mounted camera, and these are the joints that can indicate where the attention of the driver is focused. The dataset also includes binary masks indicating the positions of all the seat belts within the image. The images are sourced from 542 unique videos that include different camera angles, vehicle interiors, drivers, and passengers. To demonstrate the effectiveness of our dataset, we implement a lightweight deep learning model which simultaneously predicts both the upper-body keypoint locations and seat belt locations. We achieve an F1-score of 0.773 for the keypoint detection, and an *intersection over union* (IoU) equal to 27.8%.

## II. RELATED WORK

Our approach combines two computer vision problems: human pose estimation and seat belt segmentation. This section discusses existing HPE databases and databases designed for seat belt segmentation.

### A. Human Pose Estimation

2D human pose estimation is the task of estimating a subject's pose from a 2D image. This task is particularly relevant because any image captured with a standard RGB camera will have only 2D information of the scene, and these cameras are the most widely accessible. Human pose estimation requires a rich dataset because of the wide variety of body shapes, skin tones, and possible poses that a human can display. Furthermore, factors such as illumination and occlusion can affect the amount of information available to any HPE model in naturalistic settings. Some human poses necessarily occlude certain parts of the body from an observer (i.e. a camera). For example, a person facing to the side will have only one side of their body visible. Thus, any HPE model must be trained on a sufficiently diverse set of pose images to estimate the pose of a human subject. We make a distinction between single-person and multi-person HPE datasets. In single-person pose estimation, each image features exactly one human subject, so the model must detect the keypoints of that individual. Multi-person HPE is a more difficult task, because the model must not only identify each human subject in the image, but also uniquely assign each detected keypoint to the correct human subject. This section discusses HPE inside and outside of a vehicle. We reserve the discussion of the datasets that provide both keypoint and seat belt annotations for Section II-C.

While our main interest is in HPE datasets that feature in-vehicle images, most corpora include human poses in different contexts. Chen *et al.* [4] compare existing HPE databases in an extensive survey of 2D HPE databases. Three of the most popular databases are the MPII dataset [1], the Leeds Sports Pose dataset [16] [15], and the MSCOCO dataset [24]. These are large datasets characterized by a diverse range of backgrounds, activities, and subjects.

Our work focuses on HPE within the cabin of a vehicle, so datasets containing in-vehicle images with labeled keypoints are especially relevant. However, the number of datasets available for the specific in-vehicle domain is less than for the general HPE domain. Yoo *et al.* [34] compiled a survey of in-vehicle HPE datasets. The *driver pose estimation* (DriPE) dataset is a 2D HPE dataset collected in a naturalistic driving setting, with 19 different subjects driving around a city or closed course. Because the driving trials are not simulated, the data exhibit natural illumination changes commonly observed during motor vehicle travel. Drive&Act [25] is a large publicly available dataset that includes 3D keypoints and action labels for videos of subjects partaking in distracted driving behaviors in a parked car. Nighttime driving often yields low-light conditions in the vehicle cabin, which poses difficulty for cameras that rely on visible light. Some approaches use time-of-flight cameras to determine the distance from an object to the camera. This depth data can provide the 3D location of objects within an image, providing the resources to explore the task of 3D human pose estimation. In the Drive&Act database [25], each scene has six different synchronized camera views from different positions within the vehicle. However, this dataset does not include the location of the seat belt. Another dataset that serves as a resource for 3D HPE is the SVIRO database [8]. This corpus is a synthetic dataset featuring 3D-models of passengers and objects in the rear seat of a vehicle. The synthetic generation process allows for a variety of textures and backgrounds to be implemented in the vehicle as opposed to a single vehicle for some naturalistic datasets. As this dataset focuses on rear-seat occupancy, there are no samples containing the driver or front-seat passengers. The corpus does not include seat belt annotations.

### B. Seat Belt Segmentation

Computer vision problems involving seat belts include detection and segmentation tasks. It also includes formulations that aim to classify whether the seat belt is correctly worn in an image. Among these problems, a further distinction lies in whether the images are sourced from within the vehicle, as in our approach, or if the images come from traffic cameras outside of the vehicle.

Our approach involves in-vehicle images, which often yield higher-quality images for seat belt detection because of the subjects' closer proximity to the camera. As with in-vehicle pose estimation, few publicly available datasets exist that target in-vehicle seat belt detection. Some publicly available datasets exist from Roboflow, but these include bounding boxes for seat belt detection rather than pixel-wise segmentation masks [9] [29]. Jha *et al.* [14] used synthetic data to surmount this data deficit, augmenting a small annotated dataset with a large number of procedurally generated synthetic seat belt images.

Studies using traffic cameras outside of the vehicle have useful applications in surveillance and seat belt compliance enforcement. Common approaches include using edge-detection algorithms to detect the seat belt's characteristic

TABLE I

TABLE OF EXISTING HPE AND SEAT BELT DATASETS.

| Database | Context | Pose | Seat belt | Number of Samples | Publicly Available |
|---|---|---|---|---|---|
| MPII [1] | Out-of-vehicle | 16 keypoints | n/a | 25k Images | Yes |
| DriPE [11] | Naturalistic driving | Yes | No | 10k images | Yes |
| NADS-Net [6] | Naturalistic driving | Yes | Yes | unknown | No |
| Kim *et al.* [18] | Naturalistic driving | Yes | Yes | unknown | No |
| MSP-DISK (ours) | Naturalistic driving | 8 keypoints | Yes | 7,704 images | Yes |

edges [12], [35]. Other studies have used deep learning approaches such as *convolutional neural networks* (CNNs) to learn a visual representation of the seat belt [5], [10], [36].

### C. Combined HPE and Seat Belt Detection

Recent studies have combined the problems of in-vehicle HPE and seat belt detection to jointly predict both using a single model. Chun *et al.* [6] proposed a novel model architecture referred to as NADS-Net. This deep learning network is a lightweight CNN-based model using a feature pyramid network for keypoint and seat belt detection. The dataset contains videos of 100 drivers both in stationary vehicles and while driving, with pose and seat belt annotations for every frame. Kim *et al.* [18] used time-of-flight cameras combined with RGB cameras to efficiently annotate a dataset for 3D HPE, combining 3D body keypoint detection, seat belt segmentation, and seat belt correctness classification in a single task. In both datasets, all participants were recorded driving the same vehicle equipped with the necessary cameras and sensors. To increase the variety in vehicle interiors and seat belt colors/textures within the dataset, our proposed dataset contains samples from various vehicles, each with different seat color/texture, seat belt color/texture, camera placement, and camera quality. This increased diversity helps to prevent any model from overfitting to any particular vehicle interior. This will create a more robust model for in-vehicle monitoring systems. Additionally, our dataset will be released publicly. Table I summarizes the attributes of the most relevant datasets to ours.

## III. THE MSP-DISK CORPUS

### A. Data Collection

The dataset consists of videos collected from online video-sharing websites showing the interior of a car while a driver is operating a vehicle. These videos feature a variety of subjects, vehicles, illuminations, angles, and settings. We searched video-sharing websites with keywords related to rideshare applications, driving instruction, in-vehicle talk shows, and vlogs. We collected a total of 542 videos, from which we assigned 290 videos to the train set, 81 videos to the development set, and 171 videos to the test set. Table II shows the partitions proposed for this corpus.

Many of the videos contain sequences with different camera angles. The position of the camera within the vehicle may affect whether a part of a subject's body is visible or occluded in a given frame, thus, affecting the information available for model prediction. Figure 1 shows an example of two alternate

TABLE II

SUMMARY OF THE PARTITIONS OF THE MSP-DISK DATASET.

| **Partition** | **Videos** | **Frames** |
|---|---|---|
| Train | 290 | 5,038 |
| Development | 81 | 1,095 |
| Test | 171 | 1,571 |
| Total | 542 | 7,704 |

camera viewpoints from the same source video. To prepare a model that is robust to multiple camera angles, we attempt to include frames from different camera angles within the same video. To achieve this goal, we sample frames at a constant rate throughout the video. The sampling rate varies with the number of unique camera angles in the video. We sample fewer images from videos with few camera points of view and sample more images from videos with multiple camera angles. This strategy ensures that the dataset is not dominated by images from any particular camera location. Then, we group frames into similar clusters using a structural similarity metric [33]. Images with a high structural similarity score are more likely to show the same camera viewpoint, so we cluster these images together. Then, we manually discard any frames from scenes that show the outside of the vehicle or do not contain any subject, retaining the remaining frames for our dataset. The train set has 5,038 frames, the development set has 1,095 frames, and the test set has 1,571 frames.

### B. Data Annotation

The annotations for each frame include the pixel-wise locations of 8 joints (right/left wrist, right/left elbow, right/left shoulder, neck, and head) and a binary mask indicating the location of the seat belt. To obtain ground truth for body joint locations, we used OpenPose [3] to obtain predictions for body joints on each frame. Then, we manually correct the predictions from OpenPose and reject lower-body joint keypoints. To reduce the complexity of the keypoint detection in low-quality images, we combine all keypoints from the head area (nose, eyes, and ears) into a single keypoint located at the center of the head.

For the seat belt, we first assume that all seat belt pixels will be colored similarly within an image despite variations due to illumination differences. We use an unsupervised image segmentation approach proposed by Kanezaki [17] to group regions in the image with similar pixel colors. Then, annotators manually selected the region corresponding to the seat belt and corrected any errors in the algorithm's segmentation.

(a) Driver-side view

(b) Passenger-side view

Fig. 1. Example of different camera viewpoints within the same video. Alternative camera angles provide more variety in the data. Subject faces are blurred for anonymity in this publication.

## IV. LIGHTWEIGHT DETECTION MODEL

We demonstrate the potential use of this corpus by training a lightweight model for keypoint detection and seat belt segmentation. The motivation for building a lightweight model that performs both tasks at the same time is to reduce the computation needed to run a DMS model in the car, which is important given all the other electronic components competing for resources in the vehicle. Furthermore, we expect that simultaneously predicting keypoint and seat belt locations have advantages. When the seat belt is correctly worn, its position is relatively fixed with respect to the location of the subject's shoulders and head. Because of this observation, we hope that a model which predicts keypoint locations (including shoulders) should have an advantage when identifying the location of a seat belt within an image. If the shoulder positions are known, the most likely place that a seat belt exists is the region bounded by one shoulder and the lower part of the body.

To implement the model, we use a U-Net architecture [31] with four upsampling/downsampling layers. Figure 2 shows the model. The input to the model is a single 256x144 image. In each downsampling layer, each dimension of the image is reduced by a factor of two. Then, the model has a series of upsampling steps. In each step, we increase the image dimensions by a factor of two. The model has residual connections between the corresponding downsampling and upsampling layers. We concatenate the two same-sized images, combining the information from the downsampling stage with the upsampling stage in each step.

In many models, the output of keypoint prediction is the pixel-wise (x, y) location of the keypoint within the image. Seat belt detection is a segmentation problem, so the

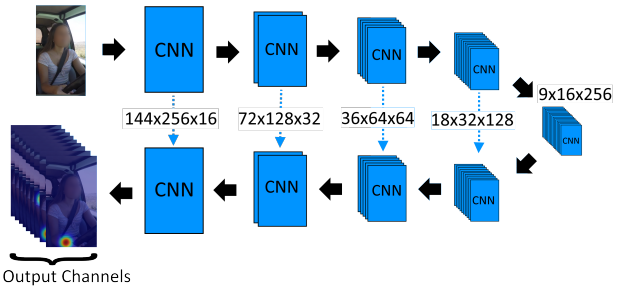


Fig. 2. The lightweight model used to simultaneously predict body keypoints and segment seat belt location. The approach is implemented with the U-Net model [31]. The outputs of the model are nine channel masks with the predicted locations for the eight body keypoints and the seat belt.

output must be an image with pixel values indicating model confidence that a seat belt exists in that pixel. Therefore, we frame the keypoint detection problem as a segmentation problem, increasing the level of similarity between the two tasks. Instead of detecting a single pixel location for each keypoint, our model predicts a region of confidence for the location of each joint. If the prediction is correct, the center of this prediction region should be close to the ground truth pixel location for the keypoint. Our formulation creates one channel image for each of the predictions of the eight joints, and one channel for the seat belt. Therefore, the output of the model is a 9-channel image, with the first 8 channels representing the model predictions for the 8 upper-body keypoints, and the 9th channel representing the model's seat belt predictions. Each channel is a grayscale image with pixel values ranging between 0 and 1. Higher pixel values indicate higher model confidence that the corresponding keypoint or seat belt is located in that pixel.

Our model has 1,947,993 parameters. The number of parameters is significantly lower than other popular computer vision algorithms. As a reference, OpenPose [3] has 21,979,692 parameters.

### A. Training Procedure

To prepare our model for predicting a confidence region near each keypoint, we pre-process the training data. We first crop each training image into one single-person image for each subject within the image. First, we select a region of the image that is 1.5x the width of the subject in the image. The subject's width is defined as the distance between the leftmost and rightmost keypoints. We then randomly shift the region left or right to prevent the central keypoints (i.e. head, neck) from being located in the same position within every training image. This step discourages the model from learning keypoint positions from absolute location within an image, rather than by identifying features. The height of this region is defined to maintain a 16:9 height:width ratio for the crop, where the subject's head is similarly randomly placed along the vertical axis. Finally, we resize the region to 256x144 pixels, representing one single-person training sample.

For each ground truth keypoint in the train set, we generate

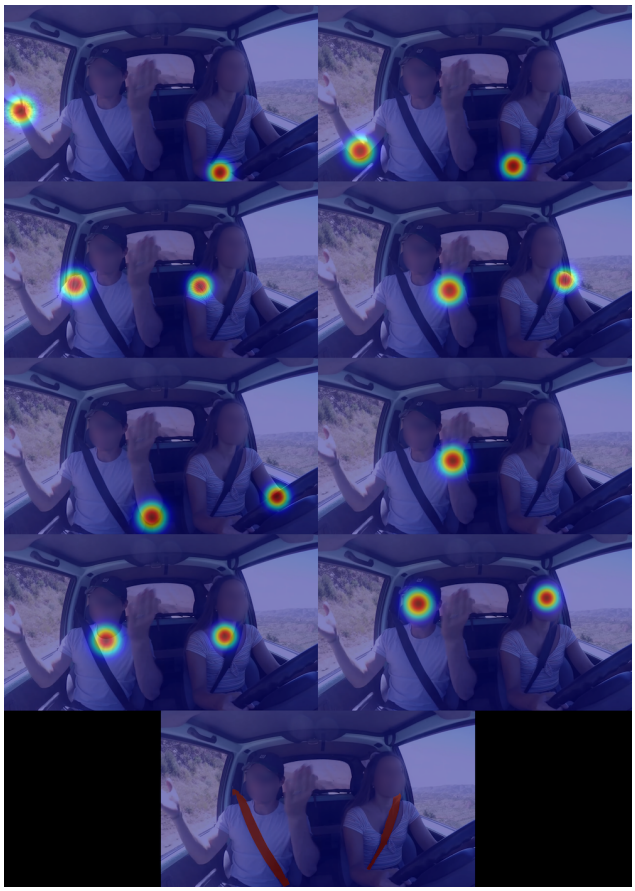

Fig. 3. Eight keypoint annotation heatmaps and seat belt annotation overlaid on a test image.

a 256x144 mask image with a 2D Gaussian map centered around the point using equation 1, where $x$ and $y$ represent each pixel in the image, and $x_0$ and $y_0$ are the ground truth keypoint locations at the center of the Gaussian distribution. $\sigma_X$ and $\sigma_Y$ are the standard deviations in the horizontal and vertical directions, respectively. By setting $\sigma_X = \sigma_Y$, the distribution becomes a circle centered at the keypoint. We scale $\sigma_X$ and $\sigma_Y$ relative to the size of the subject in the image. If the head and both shoulders are visible, each $\sigma$ is set to 1/8th of the average distance from each shoulder to the head. If at least one shoulder is not visible, we use a default $\sigma$ = 20 pixels. Figure 3 shows the 9 annotation channels overlaid on a training image.

$$f(x,y) = \exp\left(-\left(\frac{(x-x_0)^2}{2\sigma_X^2} + \frac{(y-y_0)^2}{2\sigma_Y^2}\right)\right) \quad (1)$$

For training, all 8 keypoint mask images and the ground truth seat belt mask are concatenated together in a 256x144x9 tensor. We vertically stack these channels into a 256x1296 image sent to the model as the ground truth labels for each cropped training image. Therefore, the output of the model is a 256x1296 image which, when split into a 256x144x9 tensor, contains one channel for each upper-body keypoint and one channel for the seat belt predictions. We also add a copy of each training image with a horizontal flip augmentation. In the horizontally flipped image, we swap the ground truth image channels corresponding to right-side keypoints and left-side keypoints. For example, the right wrist annotation is swapped with the left wrist annotation after the horizontal flip is applied, since flipping the image changes the subject's left side to be the right side. This augmentation improves the model's ability to generalize the representation of a wrist by including all left wrist and right wrist samples under both keypoint labels.

### *B. Loss Function*

The training loss function is a weighted combination of three loss terms. One term acts on the 8 keypoint detection channels, one term acts on the seat belt detection channel, and the final term acts on all channels equally.

For the keypoint detection loss, we use a weighted implementation of the L2 loss between the predictions and the ground truth Gaussian masks. The values assigned to each pixel in both the prediction map and the ground truth mask range from 0 to 1. A value of 0 indicates a low confidence that the pixel contains the particular keypoint or seat belt, and a value of 1 signifies a high confidence in the presence of the keypoint or seat belt. The loss from pixels where the ground truth values are greater or equal to 0.5 (high pixels) is divided by the number of high pixels, and the loss from pixels where the ground truth value are less than 0.5 (low pixels) is divided by the number of low pixels. This approach prevents the magnitude of the loss from being dominated by a large number of low pixels. We calculate this loss only for the predictions on the eight keypoint channels of the model's output.

For the seat belt segmentation loss, we use a modified version of the focal loss [23] proposed by Law *et al.* [19] for object detection. This loss is calculated between the ground truth seat belt channel and the model's seat belt prediction channel.

The final loss component is an approximation of the *intersection over union* (IoU) metric as proposed by van Beers *et al.* [32]. This loss approximates the IoU metric in a way that is differentiable for gradient calculations. IoU is defined as the ratio of the area of the intersection of two regions to the area of their union. This ratio is always between 0 and 1, where 1 represents a perfect overlap between the prediction and the ground truth. We define the IoU loss to be 1 - IoU. We apply this loss function to all 9 channels of the output.

$$\mathcal{L} = \alpha * KeypointLoss + \beta * SeatBeltLoss + \gamma * IoULoss \quad (2)$$

## V. EXPERIMENTAL RESULTS

To evaluate the usefulness of our dataset, we train a model on the train set with and without the augmentation strategy of adding a flipped copy of each training sample. For all the experiments, we set $\alpha = \beta = \gamma = 3.5$ (Eq. 2). We evaluate the performance on our test set. We also evaluate the performance of OpenPose [3], which is trained on the

TABLE III

KEYPOINT AND SEAT BELT DETECTION. THE KEYPOINT DETECTION PERFORMANCE IS MEASURED WITH PRECISION AND RECALL AND F1-SCORE. THE SEAT BELT PERFORMANCE IS MEASURED WITH IOU.

| Keypoint | OpenPose | | | Ours – Without flip augmentation | | | Ours - With flip augmentation | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | Prec. | F1 | Recall | Prec. | F1 | Recall | Prec. | F1 |
| R Wrist | 0.448 | 0.438 | 0.443 | 0.356 | 0.360 | 0.358 | 0.444 | 0.474 | **0.459** |
| R Elbow | 0.678 | 0.663 | 0.670 | 0.621 | 0.686 | 0.652 | 0.710 | 0.780 | **0.743** |
| R Shoulder | 0.943 | 0.922 | **0.932** | 0.888 | 0.935 | 0.911 | 0.912 | 0.945 | 0.928 |
| L Shoulder | 0.948 | 0.927 | **0.937** | 0.882 | 0.930 | 0.905 | 0.917 | 0.928 | 0.922 |
| L Elbow | 0.696 | 0.681 | 0.688 | 0.691 | 0.673 | 0.682 | 0.751 | 0.695 | **0.722** |
| L Wrist | 0.436 | 0.426 | **0.431** | 0.258 | 0.356 | 0.299 | 0.799 | 0.283 | 0.418 |
| Neck | 0.976 | 0.955 | **0.965** | 0.898 | 0.959 | 0.927 | 0.921 | 0.955 | 0.938 |
| Head | 0.996 | 0.975 | **0.985** | 0.901 | 0.966 | 0.932 | 0.937 | 0.936 | 0.937 |
| Average | 0.765 | 0.748 | 0.757 | 0.687 | 0.733 | 0.709 | 0.799 | 0.749 | **0.773** |
| Seat belt [IoU] | n/a | | | 21.7% | | | **27.8%** | | |

MSCOCO dataset, for keypoint detection on our test set. We report the precision, recall, and F1-score for each of the eight keypoint to illustrate the variance in prediction difficulty for some keypoints. Because the output of the keypoint detection model is a confidence heatmap, we define a true positive prediction when the center of a prediction region is within some distance $\delta$ of the ground truth keypoint location. In our experiments, we use $\delta$ = 20 pixels. A false positive is when the center of a prediction region is greater than $\delta$ pixels from the true keypoint location, and a false negative is when no prediction center is within $\delta$ pixels of a ground truth keypoint. For seat belt detection, we report the IoU between the prediction region and the ground truth region.

Table III shows the prediction results. We highlight in bold the best F1 score for each keypoint. When we compare our lightweight model, we observe that adding flip augmentations increases performance for most keypoints on the wrists, elbows and shoulders (e.g., keypoints with R and L). For example, the F1-scores for *R Wrist* and *L Wrist* increase by over 0.1 after adding flip augmentation. The OpenPose model is better than our lightweight model trained without flip augmentation. However, we can achieve better performance than the OpenPose model once the data augmentation is used. Table III shows the average performance across all keypoints. Our model trained with flip augmentations achieves the highest overall F1 score for keypoint prediction. Notice that OpenPose is a large model trained with a large MSCOCO dataset. The implementation of OpenPose that we use [26] has 21,979,692 parameters, which is over 11 times the number of parameters in our model (1,947,993). Therefore, it is remarkable that we can achieve competitive performance when using the proposed lightweight model.

For all models, the performances for the *R Wrist* and *L Wrist* keypoints are lower than the keypoints for *Shoulder*, *Neck*, and *Head*. These results suggest that the wrists are more difficult to detect regardless of what data each model was trained on. The wrists can take on a variety of positions within an image depending on the position of the subject's hands, whereas the shoulders, neck, and head are relatively stationary with respect to the subject's body and the vehicle seat.

OpenPose does not predict seat belt locations. Therefore, we compare the performance gains achieved by adding flip annotations using our lightweight model. The seat belt IoU increases by 6% (absolute) after adding the flip augmentation.

## VI. CONCLUSIONS

We presented the MSP-DISK corpus, which is a diverse, publicly available dataset for in-vehicle keypoint and seat belt detection. This dataset contains naturalistic images from multiple drivers and passengers in a wide range of vehicle settings, providing a large variability between samples. It includes annotations for the wrists, elbows, shoulders, neck, and head of the driver and all passengers, as well as the location of the seat belt for all subjects. We show that this dataset is sufficiently large to train a model for the task of keypoint detection and seat belt segmentation. The results show similar keypoint detection results to the one obtained with the OpenPose model, which is a large architecture with over 11 times the number of parameters of our lightweight model. The proposed model can simultaneously predict seat belt locations, which makes this implementation an effective DMS solution. We hope this dataset inspires further advances in the field of intelligent driver monitoring systems for the safety of drivers, passengers, and pedestrians.

Further research for collecting effective in-vehicle datasets includes crowdsourcing efforts, where individuals can opt-in to submit videos filmed within their vehicle. This research direction can provide a large variety of subjects and vehicles to improve the coverage of the database. By supplying participants with deep cameras (e.g., time-of-flight solutions), depth information can be included in the dataset. Training on the depth modality can help improve performance in low-light conditions [13].

## REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus, OH, USA, June 2014, pp. 3686–3693.

[2] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *European Conference on Computer Vision (ECCV 2020)*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Glasgow, UK: Springer Berlin Heidelberg, August 2020, vol. 12348, pp. 455–472.

[3] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, January 2021.

[4] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2D human pose estimation: A survey," *Multimedia Systems*, November 2022.

[5] Y. Chen, G. Tao, H. Ren, X. Lin, and L. Zhang, "Accurate seat belt detection in road surveillance images based on CNN and SVM," *Neurocomputing*, vol. 274, no. 24, pp. 80–87, January 2018.

[6] S. Chun, N. Ghalehjegh, J. Choi, C. Schwarz, J. Gaspar, D. McGehee, and S. Baek, "NADS-Net: A nimble architecture for driver and seat belt detection via convolutional neural networks," in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW 2019)*, Seoul, Republic of Korea, October 2019, pp. 2413–2421.

[7] Code of Federal Regulations, "Â§ 571.208 standard no. 208; occupant crash protection," 2021, retrieved May 7th, 2023. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2021-title49-vol6/pdf/CFR-2021-title49-vol6-sec571-208.pdf

[8] S. D. Da Cruz, O. Wasenmüller, H.-P. Beise, T. Stifter, and D. Stricker, "SVIRO: Synthetic vehicle interior rear seat occupancy dataset and benchmark," in *IEEE Winter Conference on Applications of Computer Vision (WACV 2020)*, Snowmass, CO, USA, March 2020, pp. 962–971.

[9] dataset, "seatbelt dataset," https://universe.roboflow.com/dataset-9xayt/seatbelt-0lhjh , may 2022, visited on 2023-05-19. [Online]. Available: https://universe.roboflow.com/dataset-9xayt/seatbelt-0lhjh

[10] A. Elihos, B. Alkan, B. Balci, and Y. Artan, "Comparison of image classification and object detection for passenger seat belt violation detection using NIR & RGB surveillance camera images," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2018)*, Auckland, New Zealand, November 2018, pp. 1–6.

[11] R. Guesdon, C. Crispim-Junior, and L. Tougne, "DriPE: A dataset for human pose estimation in real-world driving settings," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021)*, Montreal, BC, Canada, April 2021, pp. 2865–2874.

[12] H. Guo, H. Lin, S. Zhang, and S. Li, "Image-based seat belt detection," in *IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Beijing, China, July 2011, pp. 161–164.

[13] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8063–8076, July 2022.

[14] S. Jha, I. Brooks, S. Ray, R. Narasimha, N. Al-Dhahir, and C. Busso, "Seatbelt segmentation using synthetic images," in *IEEE Intelligent Vehicles Symposium (IV 2023)*, vol. To appear, Anchorage, AK, USA, June 2023.

[15] D. Johnson and M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *IEEE Conference on Intelligent Transportation Systems (ITSC 2011)*, Washington, DC, USA, October 2011, pp. 1609–1615.

[16] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference (BMVC 2010)*, Aberystwyth, UK, August-September 2010, pp. 1–11.

[17] A. Kanezaki, "Unsupervised image segmentation by backpropagation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, 2018, April 2018, pp. 1543–1547.

[18] G. Kim, H. Kim, J. Kihoon Kim, S.-S. Cho, Y.-H. Park, and S.-J. Kang, "Integrated in-vehicle monitoring system using 3D human pose estimation and seat belt segmentation," in *AAAI 2022 workshop AI for Transportation*, Vancouver, BC, Canada, February-March 2022, pp. 1–8.

[19] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *European Conference on Computer Vision (ECCV 2018)*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germany: Springer Berlin Heidelberg, September 2018, vol. 11218, pp. 765–781.

[20] T. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, "Multiple scale faster-RCNN approach to driver?s cell-phone usage and hands on steering wheel detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2016)*, Las Vegas, NV, USA, June 2016, pp. 46–53.

[21] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013, pp. 1–6.

[22] ——, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, February 2020.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV 2014)*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Zurich, Switzerland: Springer Berlin Heidelberg, September 2014, vol. 8693, pp. 740–755.

[25] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, Seoul, Republic of Korea, October - November 2019, pp. 2801–2810.

[26] MikeOfZen, "Yet another openpose implementation," Code available from Github, 2023, retrieved August 14th, 2023. [Online]. Available: https://github.com/MikeOfZen/Yet-Another-Openpose-Implementation

[27] NHTSA, "Lives saved in 2017 by restraint use and minimum-drinking-age laws," National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT), Washington, DC, USA, Technical Report DOT HS 812 683, March 2019. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812683

[28] ——, "Distracted driving 2020," National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT), Washington, DC, Technical Report DOT HS 813 309, May 2022. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813309

[29] Proudrun, "driving dataset," https://universe.roboflow.com/proudrun/driving-ag6pf , dec 2022, visited on 2023-05-19. [Online]. Available: https://universe.roboflow.com/proudrun/driving-ag6pf

[30] Y. Qiu, T. Misu, and C. Busso, "Unsupervised scalable multimodal driving anomaly detection," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 4, pp. 3154–3165, April 2023.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Munich, Germany: Springer Berlin Heidelberg, October 2017, vol. 9351, pp. 234–241.

[32] F. van Beers, A. Lindström, E. Okafor, and M. Wiering, "Deep neural networks with intersection over union loss for binary image segmentation," in *International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, vol. 1, Prague, Czech Republic, February 2019, pp. 438–445.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[34] J. S. Yoo and S. W. Jung, "Survey on in-vehicle datasets for human pose estimation," in *International Conference on Electronics, Information, and Communication (ICEIC 2022)*, Jeju, Republic of Korea, February 2022, pp. 1–2.

[35] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE International Conference on Computer Vision (ICCV 2013)*, Sydney, NSW, Australia, December 2013, pp. 1944–1951.

[36] B. Zhou, D. Chen, and X. Wang, "Seat belt detection using convolutional neural network BN-AlexNet," in *Intelligent Computing Theories and Application (ICIC 2017)*, ser. Lecture Notes in Computer Science, D.-S. H. V. Bevilacqua, P. Premaratne, and P. Gupta, Eds. Liverpool, UK: Springer Berlin Heidelberg, August 2017, vol. 10361, pp. 384–395.