100-Driver: A Large-Scale, Diverse Dataset for Distracted Driver Classification

Jing Wang, Wenjing Li, Fang Li, Jun Zhang, Zhongcheng Wu,
Zhun Zhong, and Nicu Sebe, *Senior Member, IEEE*

***Abstract*— Distracted driver classification (DDC) plays an important role in ensuring driving safety. Although many datasets are introduced to support the study of DDC, most of them are small in data size and are short of diversity in environmental variations. This largely limits the development of DDC since many practical problems such as the cross-modality setting cannot be fully studied. In this paper, we introduce 100-Driver, a large-scale, diverse posture-based distracted diver dataset, with more than 470K images taken by 4 cameras observing 100 drivers over 79 hours from 5 vehicles. 100-Driver involves different types of variations that closely meet real-world applications, including changes in the vehicle, person, camera view, lighting, and modality. We provide a detailed analysis of 100-Driver and present 4 settings for investigating practical problems of DDC, including the traditional setting without domain shift and 3 challenging settings (*i.e.*, cross-modality, cross-view, and cross-vehicle) with domain shifts. We conduct comprehensive experiments on these 4 settings with state-the-of-art techniques and show several insights to the future study of DDC. Our 100-Driver will be publicly available offering new opportunities to advance the development of DDC. The 100-driver dataset, source code, and evaluation protocols are available at https://100-driver.github.io.**

***Index Terms*— Distracted driver dataset, large-scale, cross-modality, cross-view, cross-vehicle.**

I. Introduction

CARS bring great convenience to humans and have become an indispensable part of daily travel. However, there are two sides to every door. Road traffic injuries became a growing concern that is estimated to be the seventh leading cause of death globally by 2030 [1]. According to statistics from the National Highway Traffic Safety Administration, nearly 25% of traffic accidents are caused by distracted drivers. Distracted driver behavior is any activity that takes the driver's attention away from the task of safe driving [2], [3], [4], such as using a cell phone, eating, or talking to people, to name a few. These distracted driver behaviors commonly exist during driving, which is prone to cause accidents and should be strictly avoided. Although some surveillance systems on roads can capture and identify certain types of distracted driver behavior, this can only be used as a punishment but it is not a precaution. In addition, the identification accuracy and identified types of distracted driver behavior are limited due to low image quality. Hence, it is important to develop an onboard monitoring system to alert drivers who are inattentive, greatly preventing traffic crashes.

Over the past decades, a lot of research has been introduced toward driving safety. We should notice that researchers in the naturalistic driving study (NDS) have provided large-scale datasets that include driving image data, such as SHRP2 [5], 400-car [6] and *etc*. However, the purpose of NDS is very different from DDC. NDS aims to understand driver and vehicle behavior by off-the-shell data while DDC focuses on recognizing dangerous behaviors in real-time. In addition, these NDS datasets can hardly be used for DDC tasks since most of them are not publicly available and the image quality is limited. We thus regard NDS and DDC as different tasks and do not directly compare with their datasets. For the DDC task, lots of techniques [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] were studied where posture-based distracted driver classification (DDC) has shown superiority in both terms of accuracy and efficiency [17], [18], [19], [20]. Therefore, a number of driver-posture-based datasets [21], [22], [23], [24] were proposed to support the study of DDC. However, these datasets are limited in one or more significant aspects, including scene variation, comprehensiveness of categories, and the number of drivers. To be specific, as listed in Table I, most of them are captured from a single camera view and only consider the daytime scene. In addition, the existing datasets are collected from one vehicle. These features largely limit the scene variations of existing datasets. On the other hand, most of the existing datasets consist of less than 10 distraction behaviors and less than 50 drivers. In real-world applications, the systems are deployed in different environments and undoubtedly will encounter various scenes, drivers, and behaviors. Thus, the insufficiency of existing datasets hampers the application to real-world scenarios and as such, it is essential to build a new dataset supporting the study of DDC.

In this paper, we introduce a large-scale, diverse dataset for DDC, which is simply named 100-Driver due to a

Manuscript received 5 May 2022; revised 19 November 2022 and 8 February 2023; accepted 6 March 2023. This work was supported in part by the Hefei Comprehensive National Science Center through the Pre-research Project on Key Technologies of Integrated Experimental Facilities of Steady High Magnetic Field and Optical Spectroscopy, in part by the High Magnetic Field Laboratory of Anhui Province, in part by the Hefei Advanced Computing Center, in part by the EU H2020 Project AI4Media under Grant 951911, and in part by the PRIN Project PREVUE Prot under Grant 2017N2RK7K. The Associate Editor for this article was X. Li. *(Jing Wang and Wenjing Li contributed equally to this work.) (Corresponding author: Wenjing Li.)*

Jing Wang, Wenjing Li, Fang Li, Jun Zhang, and Zhongcheng Wu are with the High Magnetic Field Laboratory, HFIPS, Chinese Academy of Sciences, Hefei 230031, China, also with the Department of Computer Science, University of Science and Technology of China, Hefei 230026, China, and also with the High Magnetic Field Laboratory of Anhui Province, Hefei 230031, China (e-mail: wjli007@mail.ustc.edu.cn).

Zhun Zhong and Nicu Sebe are with the Department of Information Engineering and Computer Science (DISI), University of Trento, 38123 Trento, Italy (e-mail: zhunzhong007@gmail.com; niculae.sebe@unitn.it).

Digital Object Identifier 10.1109/TITS.2023.3255923

TABLE I

Comparisons of Our 100-Driver With Existing Posture-Based Distracted Driver Classification Datasets. *: Only Data of 5 Divers Are Released. §: the Number of Drivers in Day/Night, and Some Drivers Appeared in Both Day and Night. -: Video Dataset That Has Many Highly Similar Frames and Can Not Be Compared With Image Datasets Fairly in Terms of Data Size. N/A: Corresponding Information Is Not Provided

	SEU-DP	**StateFarm**	**AUC**	**EBDD**	**3MDAD**	**dBehaviourMD**	**Turky-DD**	**100-Driver**
Year	2012	2016	2017	2018	2019	2020	2020	2022
Publicly available	✗	✓	✓	✓	✓	✓*	✗	✓
# Male/Female	10/10	N/A	22/9	13/0	38/12	27/10	N/A	70/30
# Day/Night §	N/A	26/0	31/0	13/0	40/19	N/A	N/A	65/52
# Vehicles	N/A	N/A	1	1	1	1	1	5
# Classes	4	10	10	4	16	13	10	22
# Cameras	1	1	2	1	2	1	1	4
Video duration	N/A	N/A	N/A	0.67 h	6.12 h	51 h	N/A	$\sim$ 79.34 h
Size	N/A	22,424	17,310	-	287,048	-	137,093	470,208
Avg. # Img/Behaviour	N/A	86.24	55.8	-	152.04	-	N/A	34.03
Resolution	640×480	640×480	1920 × 1080	854 ×480	640×480	1920 × 1080	640 × 480	1920 × 1080
Remark	Limited in one or more significant aspects. For example, scene variation, comprehensiveness of categories, and the number of drivers. The insufficiency of the existing dataset hampers the application to real-world scenarios.							Largest, Public, Diverse, New setting. Meet real application.

collection with 100 drivers. We make the four following contributions:

- **The largest public dataset**. 100-Driver contains more than 470K samples recorded over 79 hours. 100-Driver is the largest DDC dataset, which is 1.6× larger than the previous largest dataset (3MDAD [24]). It will also be publicly available.
- **The most diverse dataset**. 100-Driver is captured by 4 camera views observing 100 drivers from 5 vehicles. In addition, the samples are captured in both daytime and nighttime and are annotated across 22 categories. 100-Driver is more diverse than existing datasets and is more in line with real-world applications.
- **New settings**. Thanks to the large size and diversity of 100-Driver, we introduce four settings for DDC, including one traditional setting without domain bias, and three challenging but practical settings with domain bias. The latter are cross-modality, cross-view, and cross-vehicle settings that explicitly consider the scene variations in real-world applications.
- **Comprehensive experimental analysis and new insights**. We conducted extensive experiments on 100-Driver with state-of-the-art techniques. We validate the effectiveness of each technique in the introduced settings and reveal valuable insights to the study of DDC.

We hope our 100-Driver can encourage researchers to consider more challenging but practical problems in DDC and we believe the studies on 100-Driver have great potential to facilitate the development of DDC towards safe driving.

II. Related Works

In this section, we introduce the datasets for driver behavior analysis, including naturalistic driving study (NDS) and distracted driver classification (DDC) datasets. Although NDS and DDC are both designed for improving driving safety, they are different in terms of the objective. The goal of NDS is to understand driver and vehicle behavior by quantitative analysis based on off-the-shell data, such as "*the characteristic of crashes and indent*" [25], "*the characteristic of driver inattention*" [6] and *etc*. The conclusions analyzed by NDS can be used to guide the design of DDC. Instead, the purpose of DDC is to recognize the already-known dangerous behaviors in an online way to ensure real-time driving safety.

A. Naturalistic Driving Study (NDS) Datasets

Naturalistic Driving Study (NDS) [2], [3], [4] has made significant progress along with the emergence of the large-scale NDS datasets [5], [6], [25], [26], [27]. In 2006, the first large-scale NDS dataset 100-car [25] was conducted where 100 vehicles and 109 primary participants are involved, and multi-resource data like camera, GPS, and radar is captured. The video data in 100-car are captured from two in-car cameras and two out-car cameras which are annotated with specific events such as crashes, and near crashes with the purpose of understanding the driver and vehicle behavior in extreme circumstances. Through quantitative analysis, 100-car provides us with lots of important findings, for example, "Almost 80 percent of all crashes and 65 percent of all near-crashes involved the driver looking away from the forward roadway", "Drowsiness is a contributing factor in 12 percent of all crashes and 10 percent of near-crashes ". Subsequent datasets like UYANIK [27], SHRP2 [5], 400-car [6], UDRIVE [26] may be larger in terms of the number of divers or vehicles, or more diverse in road situation and data modality, their goal is still similar to 100-car. For example, the goal of the 400-car [6] dataset is to understand the driver's behaviors in normal, impaired, and safety-critical situations while UDRIVE [26] aims to obtain a better understanding of drivers' engagement in secondary task activities.

However, most of the existing NDS datasets can not be directly utilized for DDC. (a) First and foremost, existing NDS datasets are not publicly available (especially video data) for academic study [5], [6]. (b) The image quality in most NDS datasets is pretty low [5] because they are often highly compressed to meet the requirement of the storage of a large amount of data. Additionally, the videos in NDS datasets often have a very low frame rate (*e.g.,* 10 FPS), resulting in the loss of important frames. These two factors will increase the difficulty of using NDS datasets for DDC.

B. Distracted Driver Classification (DDC) Datasets

Considering the requirements of both effectiveness and efficiency in real-world applications, although driver physiological information [7], [15], [28] or vehicle kinematic signatures [8], [29] can be used for DDC, recognizing distracted drivers in a visual manner is a better choice [25]. The vision-based DDC datasets can be divided into two categories, body-part-based and posture-based datasets.

1) Body-Part-Based DDC Datasets: Body-part-based DDC datasets extract drivers' head [30], [31], [32], [33], facial (*e.g.,* face [10], [34], [35], eyes [11], [12], and mouth [36], [37]), and hand [9], [38] features to recognize several specific distraction behaviors. To be specific, driver head datasets such as DriveAHead [30], LISA-P [33], and CoHMEt [32] aim to monitor driver awareness by estimating the head position and rotations like yaw, roll, and pitch. Driver eyes datasets [11], [12] are to recognize the behavior of fatigue, sleepiness, and inattention based on the driver's eye states such as the eye blinking frequency and eye closure duration. Similarly, driver mouth datasets [36], [37] are conducted to determine the yawning behavior by analyzing the driver's mouth opening level.

Despite their low computational cost [39], the models trained on body-part-based DDC datasets have two limitations. First, they are sensitive to scene variation. For example, the models trained on the facial datasets will come to nothing if the driver just wears a mask or sunglasses [39]. Second, the models trained on body-part-based DDC datasets can recognize limited distracted behaviors. These two limitations largely restrict their subsequent real-world applications. Compared to body-part-based DDC datasets, posture-based DDC datasets try to capture the distraction behaviors by the driver's whole posture, which are more robust to variations and cover a more comprehensive set of distractions [21], [22], [24].

2) Posture-Based DDC Datasets: In recent years, a number of posture-based datasets have been released for distracted driver classification (DDC). Although SEU-UP is the first dataset for DDC, it is not publicly available and only indicates the driver information. Several years later, StateFarm [21] and AUC [22] are introduced to support the study of DDC, which however are limited in the data size. Recently, two large-scale datasets, 3MDAD [24] and Turky-DD [40], are proposed, which have more than 287K and 137K images, respectively. Nevertheless, both of them have low image quality and Turky-DD is not publicly available.

The above five datasets are image-based datasets. There are two video-based datasets presented in the community, EBDD [41] and dBehaviourMD [23]. Although dBehaviourMD includes more than 1M images, it is indeed not as diverse as 3MDAD and Turkey-DD since there are many highly similar frames in it. In addition, dBehaviourMD only releases a small portion of the data (5 of 37 drivers), largely reducing its data size. Despite the wide use of the above datasets, all of them are relatively limited in data diversity. This leads them still far from real-world scenarios and hinders the investigation of DDC. To this end, we build a large-scale, diverse dataset for DDC to narrow the gap from real-world applications. Our 100-Driver includes more than 470K images and is diverse in terms of driver, vehicle, camera view, and class. Compared to 3MDAD, our 100-driver offers {2×, 5×, 2×, and 1.3×} more {drivers, vehicles, views, and classes} respectively. A comparison of different datasets is listed in Table I.

III. The 100-Driver Dataset

In this section, we first present the data generation process of 100-Driver. Then we detail the dataset statics and analyze the dataset properties. Last, we introduce four settings for practical evaluation.

A. Dataset Generation

1) Collection Setting: During data collection, we elaborately control the diversity of raw data in terms of vehicles (5 vehicles, Mazda 3 axela, Lynk&co 03, Toyota C-HR, Hyundai X25, and Ankai A6), camera locations (4 Xiaomi-C1 cameras in front-left, front, front-right, and side-right, as shown in Figure 4), modalities (RGB and Near Infrared NIR), lighting conditions (from morning to afternoon, from spring to winter, and different weather conditions), drivers (100 participants), appearance variations (changing clothes, wearing a mask, hat, and sunglasses). The RGB modality is captured in the daytime while the NIR modality is collected in the nighttime. To ensure the appearance variations, a part of drivers (25% in daytime and 15% in nighttime) were recorded over multiple time periods, leading to substantial appearance variations, especially in clothes and lighting (see the last row of the left sub-figure in Figure 1). During collection, we equip a safety officer to give relevant instructions according to the road condition. Note that the officer is only in charge of announcing the name of the distracted behaviors from a set as listed in Table II. We record the natural reaction of each driver and no other intervention such as telling the participants how to perform such behavior and etc. And the announced sequence for each participant is random. *Each participant is informed of the risks involved in data collection and has signed a General Data Protection Regulation (GDPR) informed consent to allow the data to be publicly available for research study.*

2) Data Annotation: Following the collection setting, we initially obtain 79.34 hours of video. The overall annotation process was conducted by 20 experts. To boost the efficiency of data annotation, we first grouped the data by drivers. In addition, we aligned the start and end times for the 4 cameras of the same driver, so that we could label each pre-defined class (as listed on Table II) for all 4 cameras at once based on the timestamp. Each individual behavior is labeled with behavior class, modality type, driver ID, camera ID, vehicle ID, and scene ID. Given the labeled video clips, we conduct downsampling to generate more diverse data considering the high similarity between adjacent frames. We further remove outliers with very different content from the labeled class. *Note that, the downsampling and filtering processes lead our 100-Driver to be much more diverse and clean than the previous largest dataset, 3MDAD [24], which builds the data with the video clips directly.* In Table I, we compare the average number of images of each individual behavior and the overall data size among datasets, showing that the large-scale data of 100-Driver mainly benefited from collecting more diverse samples instead of highly similar ones. We finally produce a total of 470,208 samples to form our 100-Driver dataset. An example of the annotated samples is shown in Figure 3.

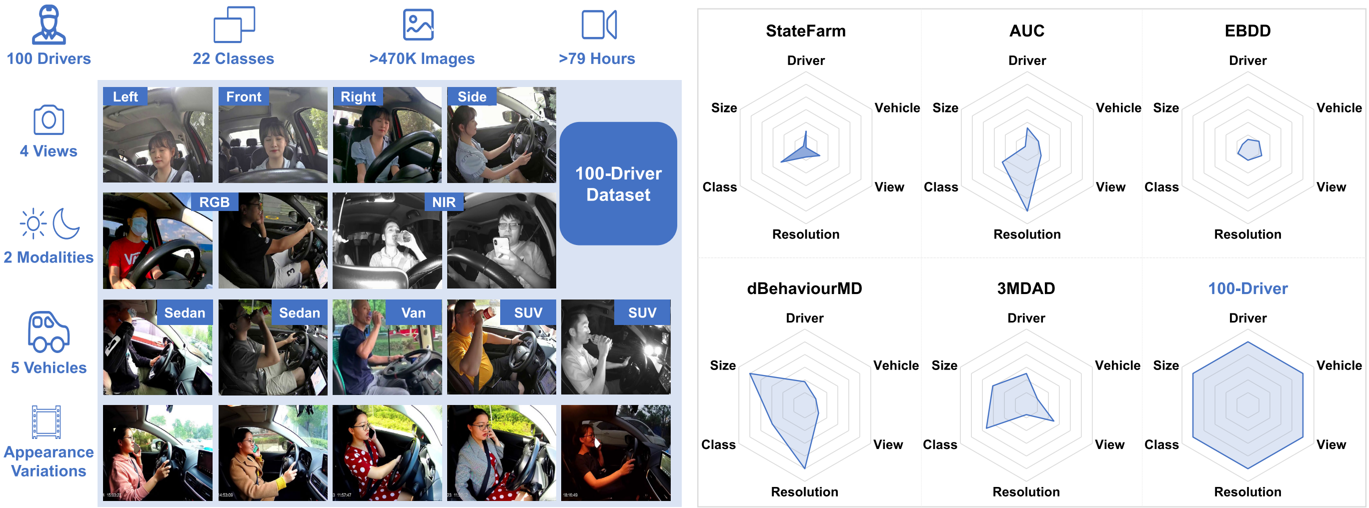


Fig. 1. **Left**: The specification of the proposed 100-Driver for distracted driver classification. **Right**: The comparisons of different datasets. For each item, the biggest value among all datasets is used to determine the unit, which is denoted as V_b. We first calculate the unit by $V_b/10$ and then normalize the value of each dataset by dividing it through the obtained unit. Each scale represents 2 units.

TABLE II
THE LIST OF DISTRACTED DRIVING BEHAVIORS IN THE 100-DRIVER DATASET

No.	Behavior
1	Normal driving
2	Sleeping
3	Yawning
4	Talk with cellphone (left)
5	Talk with cellphone (right)
6	Texting (left)
7	Texting (right)
8	Hair / makeup
9	Looking left
10	Looking right
11	Looking up
12	Looking down
13	Smoking (left)
14	Smoking (right)
15	Smoking (mouth)
16	Drinking / Eating (left)
17	Drinking / Eating (right)
18	Adjusting radio
19	Operating GPS / entertainment system
20	Reaching behind
21	Hands off the steering wheel
22	Talking to passengers

B. Dataset Description

1) Overview: 100-Driver contains 470,208 samples, which are collected by 4 camera views and belong to RGB and NIR modalities. It involves 21 types of distracted behaviors and 1 normal behavior captured from 100 drivers in 5 vehicles. The detailed classes can be found in Figure 2(c) and (d).

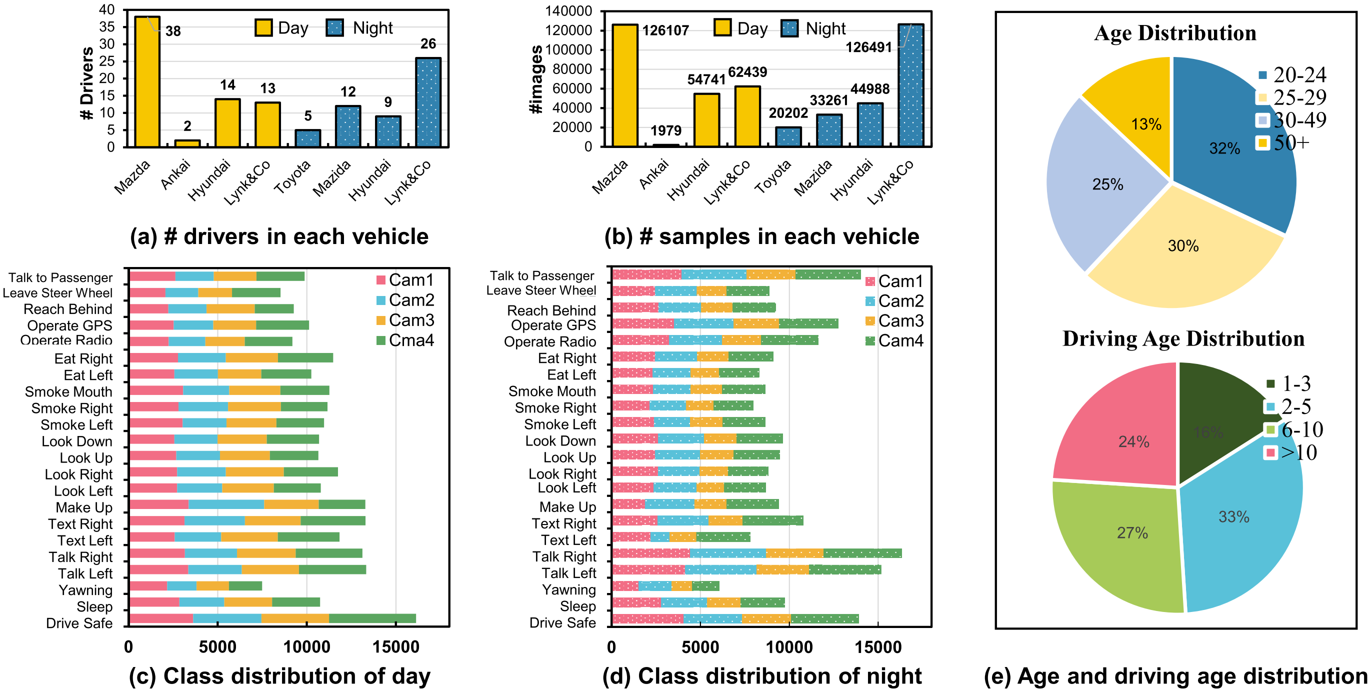
2) Data Statics: In Figure 2, we provide the detailed distributions of 100-Driver. We can make the following conclusions. *First*, the number of samples and the number of drivers are roughly balanced between daytime and nighttime. Specifically, there are 65 and 52 drivers recorded during daytime and nighttime, respectively, where 17 drivers participated in both events. In total, there are 245,266 RGB images captured in the daytime and 224,942 NIR samples captured in the nighttime. *Second*, the data distribution of each vehicle is different. In detail, three vehicles collect the data in both daytime and nighttime while the other two only collect the data in daytime or nighttime. Besides, the "Mazda" and "Lynk&Co" are the two vehicles collected with the most number of samples where the "Mazda" mainly focuses on the daytime while the "Lynk&Co" is the opposite. This is because, in our original intention, we only considered the balance between day and night, *i.e.*, the number of samples and the number of drivers are roughly balanced between daytime and nighttime. During collection, the available long-term vehicles were Mazda, Hyundai, and Lynk&Co. To increase data diversity, we temporarily used the other two vehicles (Ankai and Toyota) and asked new participants to collect the data. This collection strategy led to a slight unbalance in our data. *Third*, the class distributions are relatively balanced regarding both cameras and modalities. To be specific, the class distributions are similar between daytime and nighttime. And the number of samples of the same class is approximate for each camera. The Yawning class has the smallest number of samples because the duration of individual behavior of this class during collection is shorter than other classes. Lastly, as shown in Figure 2 (e), our dataset covers diverse participants in terms of age group and driving experience.

C. Data Properties

Our 100-Driver gains benefit in two aspects, *i.e.*, scale, and diversity, which are explained below.

I: 100-Driver has scale advantage in terms of the number of samples and the number of drivers.

- **100-Driver has the largest number of samples.** As shown in Table I, 100-Driver contains 470K images,



(a) # drivers in each vehicle

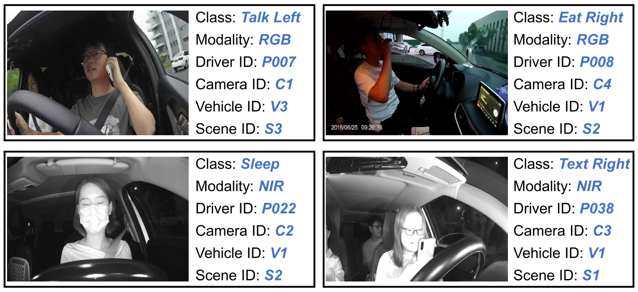
(b) # samples in each vehicle

(c) Class distribution of day

(d) Class distribution of night

(e) Age and driving age distribution

Fig. 2. Statistics of the 100-Driver dataset.



Class: *Talk Left*
Modality: *RGB*
Driver ID: *P007*
Camera ID: *C1*
Vehicle ID: *V3*
Scene ID: *S3*

Class: *Eat Right*
Modality: *RGB*
Driver ID: *P008*
Camera ID: *C4*
Vehicle ID: *V1*
Scene ID: *S2*

Class: *Sleep*
Modality: *NIR*
Driver ID: *P022*
Camera ID: *C2*
Vehicle ID: *V1*
Scene ID: *S2*

Class: *Text Right*
Modality: *NIR*
Driver ID: *P038*
Camera ID: *C3*
Vehicle ID: *V1*
Scene ID: *S1*

Fig. 3. The label of samples in the 100-Driver dataset.

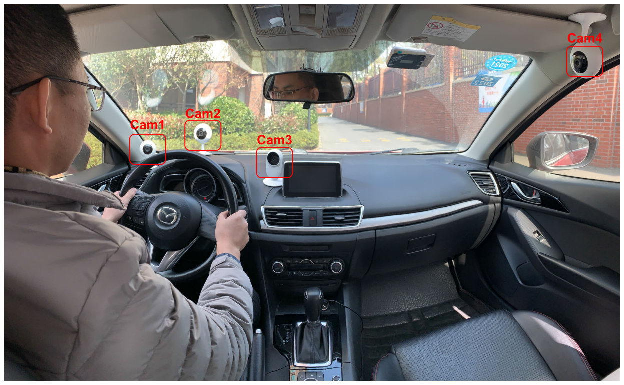


Fig. 4. The camera locations in the 100-Driver dataset. Cam1: front-left, Cam2: front, Cam3: front-right, and Cam4: side-right.

which is 1.6× larger than the previous largest dataset, 3MDAD [24].

- **100-Driver has the largest number of drivers.** 100-Driver recorders the samples of 100 drivers, which is substantially larger than existing datasets that cover fewer than 50 drivers.

II: 100-Driver has a diversity advantage in terms of distracted class, behavior style, camera view, vehicle, and person appearance.

- **100-Driver covers the most comprehensive classes.** 100-Driver considers 22 classes (21 distracted classes and 1 normal driving) while 3MDAD only considers 16 classes (15 distracted classes and 1 normal driving). The coverage of distraction behaviors is designed according to the definition of the International Road Transport Union. Examples of distracted classes and safe driving classes are shown in Figure 5. Specifically, classes like "look away" and "leave the steering wheel" are first defined compared to previous datasets.
- **100-Driver involves more diverse behaviour styles**. Since peoples have different habits during driving, they will react differently to each behavior. We invite 100 drivers of different age groups (from 20-60 years old) during collection, leading the behavior styles to be more diverse.
- **100-Driver captures samples with 4 different views.** During collection, four cameras are placed in front-left, front, front-right, and side-right views of the drivers, with two purposes. First, a multi-camera dataset can help us to learn models that are more robust to camera variations as well as enable us to evaluate the generalization ability of models to cameras. Second, the multi-camera dataset provides an opportunity to boost the system's performance by considering the contents captured by multiple cameras. For example, the front-view camera is good at capturing the facial details that are important to detect subtle activities such as sleep and yawning distractions. The side-view camera can provide a global view of an action, which is more suitable to identify behaviors with large movements, such as reaching behind.

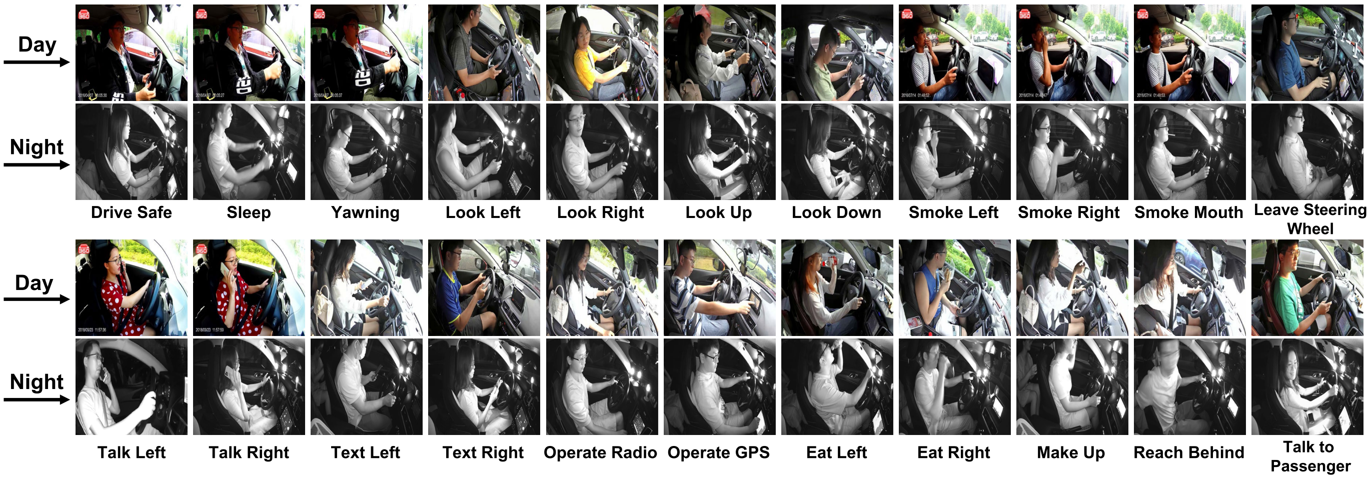


Fig. 5. Samples of the proposed dataset in side view.

- **100-Driver is collected from different vehicles.** In real-world applications, the monitoring system will be installed in different vehicles that have very different in-car scenes. To meet real-world applications as much as possible, 100-Driver is comprised of 5 vehicles, including 2 sedans, 2 SUVs, and 1 van. To our best knowledge, 100-Driver is the first dataset that considers the diversity of vehicles.
- **100-Driver includes various lighting and person appearance variations.** In our dataset, three factors lead to large person appearance variations. First, the data are collected under different weather conditions (sunny, rainy), periods (morning, noon, afternoon), and seasons (summer and winter). This leads to lighting and clothes variations. Second, some participants are asked to change their clothes, wearing masks, sunglasses, and hats, further enlarging the appearance variations. Third, more drivers also lead to clothes variations since the clothes worn by different drivers are very different.

In summary, 100-Driver is a large-scale, diverse dataset that explicitly considers the important factors in real-world applications. This enables us to study more practical problems in DDC as presented in the next section.

D. Evaluation Protocol

In previous datasets, they generally assume that the training and testing sets have the same distribution. That is, the training and testing sets are collected under the same environments, including camera views, vehicles, and modalities. However, in real-world applications, the deployed environments vary significantly. Therefore, the trained model inevitably needs to evaluate the data collected from environments that are very different from the training ones. Considering the above fact, the traditional setting ignoring the domain bias is not always practical, and it is essential to evaluate settings that consider the variations caused by changes in camera views, vehicles, and modalities. However, as discussed before, existing datasets commonly are collected under a single environment (*e.g.*, with 1 camera, 1 vehicle, and 1 modality) and thus can not be used to evaluate the challenging settings with domain bias. Thanks to the high diversity of our 100-Driver, we are able to achieve this goal and thus introduce four settings to narrow the gap from the practical deploying scenarios.

TABLE III

Description of Different Settings. i, j: Refers to Driver IDs. c: Camera ID. t: Refers to Vehicle Type. m: Refers to Vehicle ID

Setting	Train	Test
Traditional	Driver$_1$,...,Driver$_i$	Driver$_{i+1}$,...,Driver$_j$
Cross-view	Day-Cam$_c$	Day-Cam$_{\{\text{NOT } c\}}$
Cross-modality	Day-Cam$_c$	Night-Cam$_c$
Cross-vehicle	Day-Vehicle$_{\{t\}}$ Day-Vehicle$_{\{m\}}$	Day-Vehicle$_{\{\text{NOT } t\}}$ Day-Vehicle$_{\{\text{NOT } m\}}$

- **Traditional Setting.** In this setting, the training and testing sets are captured from the same camera views, modalities, and vehicles. The domain bias between training and testing sets is very small.
- **Cross-camera Setting.** In this setting, the training and testing sets are collected from different cameras while the modalities and vehicles are the same. The domain bias is mainly caused by camera variations.
- **Cross-modality Setting.** Similar to the cross-camera setting, in this setting, the training and testing sets are collected from different modalities while the camera views and vehicles are the same. The domain bias is mainly caused by the modality difference.
- **Cross-vehicle Setting.** Cross-vehicle setting includes cross-vehicle-type and cross-individual-vehicle settings. The training and testing sets are collected from different vehicle models and different vehicles for cross-vehicle-type and cross-individual-vehicle, respectively. And the camera views and modalities are the same for both cross-vehicle-type and cross-individual-vehicle. The domain bias is mainly raised by vehicle type and vehicle changes.

IV. Experiments

In this section, we conduct extensive experiments on the proposed 100-Driver dataset.

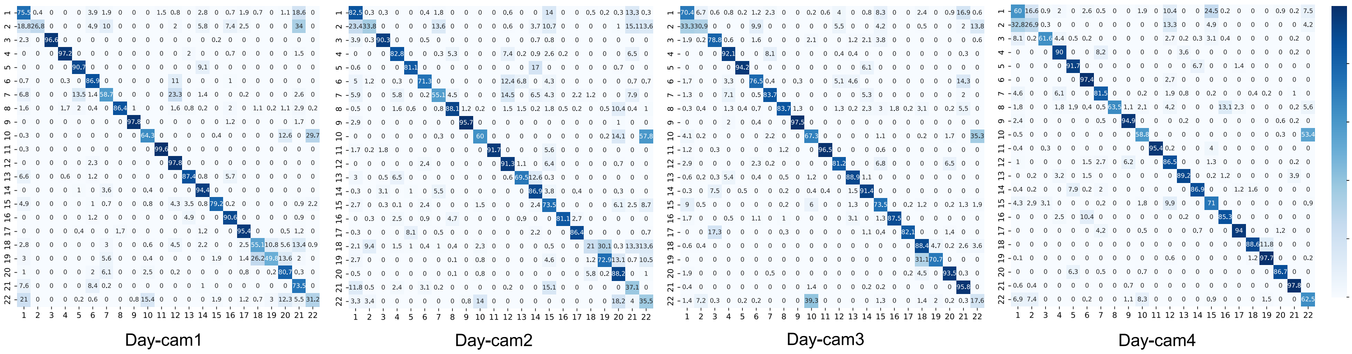


Fig. 6. The confusion matrix of EfficientB0 on four cameras. And the best view is in zoom.

A. Experimental Setting

1) Baseline: Since Distracted driver classification (DDC) is a classification problem, we use the cross-entropy loss to train the model and regard this approach as the baseline.

2) Backbones: We select 6 popular networks as the backbones, including ResNet-50 [42], MobileNetV3-large [43], ShuffleNetV2-1-0 [44], SqueezeNet1-0 [45], EfficientNet-B0 [46], and GhostNet-1.0 [47]. *Note that, in this paper, we do not aim to compare the performance of different backbones. Instead, we hope to find common phenomena that are important to DDC.*

3) Evaluation: We evaluate the baseline method on the 4 settings introduced in Section III-D. *For traditional setting*, we split the data by driver, where {47, 6, 12} and {37, 5, 10} drivers are divided into {train, val and test} sets for daytime and nighttime, respectively. *For cross-camera setting*, we use the data of one camera to train the model and use it to evaluate the testing data of other cameras. *For cross-modality setting,* we adopt the data of one modality to train the model and use it to evaluate the testing data of another modality. *For cross-vehicle setting,* we split the data by vehicles, where the training sets for the day are comprised of the data recorded in Mazda. The data in Mazda are divided by driver where {33, 5} drivers are for training and testing. The data collected by {Hyundai, Ankai, Lynk&Co} are for testing. For all settings, we select the model that achieves the best accuracy on the validation set and report the accuracy on the testing set.

4) Implementation Details: For baseline models, we adopt SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The batch size is set to 64. All backbones are pretrained with ImageNet [48]. The learning rate is initialized to 0.01 and reduced by a factor of 10 at 40 and 60 epochs. The inputs are resized to 224 × 224. We use random crop and random erasing [49] for data augmentation. The overall training epoch is 100.

B. Results on Traditional Setting

We first evaluate the traditional setting with different backbones in Table IV. We find that the models commonly have higher accuracies when training and testing on Camera 1 (front-left) or Camera 4 (side-right), regardless of the backbone and modality. We also provide the results of precision and recall and find a similar phenomenon as the accuracy metric. This indicates that, in real-world applications, we can suggest/enforce the drivers to install the cameras at the front-left and side-right. We also find that, with the same camera and backbone, the model generally achieves higher accuracies in the daytime. For instance, when using ResNet50 as the backbone and evaluating on Camera 4, the model achieves 77.3% accuracy under daytime while obtains 74.1% accuracy under nighttime. This phenomenon is reasonable since poor lighting at nighttime will increase the difficulty of recognition.

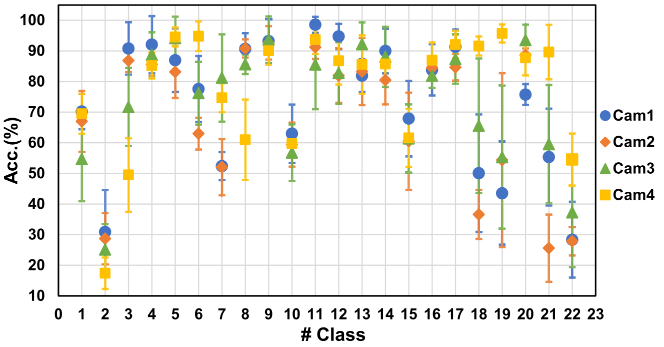


Fig. 7. The accuracy of each class. For each class, we compute the mean and standard deviation of the accuracy of all the models. And the best view is in zoom.

To further investigate the challenges of 100-Driver, we take a closer look at the accuracy of each class in Figure 7 and the confusion matrix in Figure 6. We can make the following observations. First, the accuracies are imbalanced for different classes, indicating that the difficulties of each class are not consistent. For instance, the classes "sleeping"(#2), "hair / make up "(#7), "looking up" (#10), and "talking to passengers"(#22) are the most difficult behaviors, regardless of the cameras and models. Therefore, future studies may consider using data re-sampling [50] or class re-weighting [51], [52] techniques to improve the overall performance. Second, different cameras are good at capturing different distractions. For example, Camera 1 shows superiority in identifying the "yawning" (#3) activity while Camera 3 and Camera 4 perform poorly in recognizing such class. Camera 4 can well classify the "operating GPS / entertainment system"(#19) whereas the other three cameras can hardly distinguish such distractions. This phenomenon indicates that one can use a multi-camera fusion strategy to take advantage of cameras installed at different views. For example, during the fusion of predictions produced by multiple cameras, the weights of each class will be set according to the superiority of each camera. Third, the

TABLE IV

THE RESULTS OF THE TRADITIONAL SETTING. D AND N INDICATE DAY AND NIGHT, RESPECTIVELY. THE NUMBER FOLLOWING D AND N INDICATES THE CAMERA ID, WHERE "ALL" REPRESENTS ALL CAMERAS. ACC, PRE AND REC REPRESENT ACCURACY (%), PRECISION (%) AND RECALL (%), RESPECTIVELY

Model	D-All			D1			D2			D3			D4			N-All			N1			N2			N3			N4		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
Resnet50	73.9	74.4	72.7	71.5	72.8	70.7	68.0	70.6	65.7	69.9	71.9	67.8	77.3	80.3	75.7	74.3	74.3	73.5	66.3	64.7	57.4	55.7	51.5	40.4	53.3	54.7	47.5	74.1	73.8	69.5
MobileNetV3	76.4	77.0	75.5	74.1	72.9	74.0	71.7	69.6	69.8	76.0	74.8	75.0	77.2	76.5	76.7	71.4	71.3	70.7	70.1	72.2	70.6	67.9	68.3	66.1	67.0	65.1	65.3	75.1	78.1	74.3
ShuffleNetV2	74.4	74.5	73.3	70.0	69.8	69.0	64.6	63.3	63.4	67.0	66.3	67.2	74.7	72.7	74.3	69.9	70.7	68.8	74.8	76.2	74.7	62.5	63.1	59.3	64.4	65.8	62.7	72.6	75.1	71.8
SqueezeNet	72.3	72.0	71.6	75.0	73.6	73.9	72.3	72.4	70.1	79.6	78.5	79.2	82.5	80.4	82.2	70.5	73.2	69.4	75.2	74.9	74.7	65.9	67.1	63.7	66.7	68.6	66.3	77.1	77.2	76.9
EfficientNetB0	79.0	79.2	78.1	79.8	79.6	78.9	72.3	72.2	72.5	77.2	75.3	77.4	78.6	78.9	78.0	74.1	74.6	73.7	72.3	73.2	71.9	67.9	68.0	65.4	64.6	65.6	63.2	74.6	75.5	73.2
GhostNetV1	72.3	72.4	71.2	70.7	70.2	70.2	68.4	65.7	66.1	72.9	71.0	71.9	75.2	77.0	74.7	66.4	65.8	65.5	69.8	72.1	71.0	61.8	57.6	58.1	63.5	64.0	61.1	72.2	77.5	71.7

TABLE V

ACCURACY (%) OF MULTI-CAMERA FUSION ON DAYTIME. D INDICATES DAY. THE NUMBER FOLLOWING D INDICATES THE CAMERA ID. THE FIGURES IN BRACKETS INDICATE HOW MUCH THE ACCURACY OF n-CAMERA COMBINATION INCREASES OR DECREASES COMPARED TO THE BEST RESULT OF INDIVIDUAL OR $n-1$ COMBINED SET WHERE ↑ AND ↓ REPRESENT INCREASE AND DECREASE, RESPECTIVELY

Model	D1 & D2	D1 & D3	D1 & D4	D2 & D3	D2 & D4	D3 & D4	D1&D2&D3	D1&D2&D4	D1&D3&D4	D2&D3&D4	D1&D2&D3&D4
ResNet50	73.8 (↑2.3)	74.7 (↑3.2)	82.5 (↑5.2)	72.8 (↑2.9)	78.7 (↑1.4)	80.8 (↑3.5)	76.8 (↑2.1)	82.5 (↑0)	83.3 (↑0.8)	83.0 (↑2.2)	84.4 (↑1.1)
MobileNetV3	77.1 (↑3.0)	81.9 (↑5.9)	83.8 (↑6.6)	77.9 (↑1.9)	82.6 (↑5.4)	82.2 (↑5.0)	83.5 (↑1.6)	86.1 (↑2.3)	84.8 (↑1.0)	84.6 (↑2.0)	86.9 (↑0.8)
ShuffleNetV2	73.2 (↑3.2)	75.1 (↑5.1)	79.9 (↑5.2)	72.2 (↑5.2)	76.1 (↑1.4)	76.7 (↑2.0)	77.2 (↑2.1)	78.4 (↓1.5)	79.7 (↓0.2)	77.8 (↑1.1)	80.0 (↑0.3)
SqueezeNet	82.1 (↑7.1)	80.4 (↑0.8)	86.2 (↑3.7)	83.6 (↑4.0)	85.2 (↑2.7)	86.2 (↑3.7)	82.9 (↓0.7)	84.9 (↓1.3)	86.5 (↑0.3)	84.7 (↓1.5)	83.5 (↓3.0)
EfficientNetB0	81.7 (↑1.9)	83.6 (↑3.8)	87.9 (↑8.1)	80.0 (↑2.8)	83.2 (↑4.6)	82.9 (↑4.3)	84.8 (↑1.2)	87.8 (↓0.1)	86.8 (↓1.1)	84.5 (↑1.3)	86.9 (↓0.9)
GhostNetV1	74.8 (↑4.1)	78.0 (↑5.1)	83.7 (↑8.5)	75.8 (↑2.9)	80.1 (↑4.9)	82.4 (↑7.2)	82.1 (↑4.1)	85.1 (↑1.4)	85.0 (↑1.3)	83.6 (↑1.2)	86.2 (↑1.1)
Computation Cost	2×						3×				4×

accuracy of "normal driving" (#1) is not satisfied (lower than 90% in all cases). This is very dangerous in practice since recognizing distracting behaviors as normal driving should be more worried than classifying them into wrong distracted behaviors. Therefore, the researchers should take into account this weakness seriously.

1) Multi-Camera Fusion: To study the trade-off between accuracy and computational cost in the multi-camera fusion approach, we propose a simplified framework to fuse multiple cameras by averaging the prediction of each camera. An example of a two-camera fusion framework is illustrated in Figure 8. The experimental results are listed in Table V. And we obtain several observations as follows. First, two-camera fusion can consistently boost the performance regardless of the backbone model, especially on the combination of Camera 1 and Camera 4 which can achieve the best performance in most cases. For example, the improvements based on EfficientNetB0 and GhostNetV1 can reach 8.1% and 8.5%, respectively. We conjecture this is because the data captured from Camera 1 and Camera 4 can be better complementary. Second, the three-camera, as well as four-camera fusion methods, can obtain limited improvement or even cause a performance drop. For example, the combination of Camera 1, Camera 2, and Camera 4 can damage the performance in half cases, in which the accuracy is reduced by 1.5% compared to using the combination of Camera 1 and Camera 4 with ShuffleNetV2. This may be because the introduction of Camera 2 might destroy the effect of the fusion of Camera 1 and Camera 4 which have a good complementary relationship. Third, the computational cost and parameter size grow linearly with the number of cameras. Therefore, to better trade the accuracy and the complexity, we suggest combining the information from Camera 1 and Camera 4. And two directions can be further studied to reduce the computational cost. (a) It would be more effective to design some parameter-sharing models for multi-camera inputs. (b) An alternative way is to design a parallel computing solution from the hardware level.

Research Findings.

- Different camera locations produce different results where Camera 1 and Camera 4 perform better.
- The difficulties of each class are not consistent, where "Normal driving", "Sleep", "Smooking" are harder than other classes.
- Two-camera fusion can boost performance.
- Three-camera and four-camera fusion obtain limited improvements.

Suggestions for Deployment.

- Preferentially installs the cameras at the front-left and side-right locations.
- To achieve higher accuracy, consider jointly enabling two views for recognition.

Future Directions.

- Improve the performance in recognizing hard classes, especially the "Normal driving".
- Effectively explore the mutual benefit of different views to obtain more robust results.

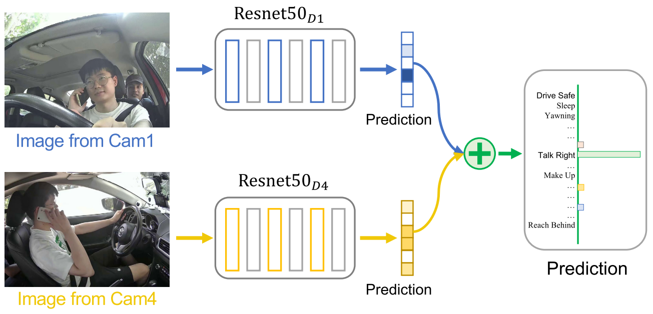


Fig. 8. A simplified two-camera fusion framework. $Resnet50_{D_i}$ denotes the model trained on Camera i.

C. Results on Cross-Domain Settings

We then conduct experiments under the cross-camera, cross-modality, and cross-vehicle settings. We use the ResNet-50

TABLE VI
Accuracy (%) of Cross-View Setting on Daytime. D Indicates Day. The Number Following D Indicates the Camera ID

Model	D1→ D2	D1→ D3	D1→ D4	D2→ D1	D2→ D3	D2→ D4	D3→ D1	D3→ D2	D3→ D4	D4→ D1	D4→ D2	D4→ D3
ResNet50	50.1	18.4	6.1	11.2	30.4	6.1	15.6	31.4	13.1	5.4	4.1	15.0
MobileNetV3	48.7	15.0	4.0	16.6	32.1	2.8	12.9	25.3	9.1	4.2	3.5	9.6
ShuffleNetV2	44.1	14.7	5.8	18.9	21.9	5.3	7.8	26.8	8.8	3.7	3.4	8.5
SqueezeNet	52.1	19.6	5.8	31.3	38.3	5.4	14.1	31.8	11.7	4.9	5.2	11.1
EfficientNetB0	51.3	17.3	5.0	20.7	27.8	4.0	10.4	28.3	9.0	5.7	3.8	9.1
GhostNetV1	48.0	13.1	6.8	20.5	24.1	4.5	12.6	25.3	11.8	3.5	4.0	8.9

TABLE VII
Accuracy (%) of Cross-Vehicle Setting. D Indicates Day. The Number Following D Indicates the Camera ID. Se Represents Sedan

Model	D1			D2			D3			D4		
	Se→Se	Se→SUV	Se→Van	Se→Se	Se→SUV	Se→Van	Se→Se	Se→SUV	Se→Van	Se→Se	Se→SUV	Se→Van
ResNet50	55.6	36.2	5.2	62.4	28.5	1.5	60.4	25.4	7.9	73.2	42.3	8.0
MobileNetV3	56.0	34.1	7.4	63.0	30.7	32.5	59.7	23.9	4.8	68.1	36.7	0.8
ShuffleNetV2	51.1	28.6	4.6	56.7	24.2	1.3	52.8	28.8	5.0	65.0	38.3	6.0
SqueezeNet	57.6	36.1	5.4	67.5	31.3	19.1	66.0	38.4	24.1	72.1	36.4	25.9
EfficientNetB0	57.7	34.1	4.1	63.2	30.3	15.9	65.6	28.1	19.4	71.5	42.5	11.3
GhostNetV1	55.4	31.7	2.1	56.7	29.5	9.3	60.7	27.3	11.8	65.6	38.9	0.25

TABLE VIII
Accuracy (%) of Cross-Vehicle Setting. D Indicates Day. The Number Following D Indicates the Camera ID. {M, H, A, L} Represent {Mazda, Hyundai, Ankai, Lynk&Co}

Model	D1				D2				D3				D4			
	M→M	M→H	M→A	M→L	M→M	M→H	M→A	M→L	M→M	M→H	M→A	M→L	M→M	M→H	M→A	M→L
ResNet50	54.9	27.7	12.3	29.6	60.5	22.8	0.8	32.6	61.9	18.9	4.1	29.8	73.9	32.5	16.8	34.0
MobileNetV3	55.1	26.7	11.0	25.9	67.8	24.1	26.3	32.9	55.1	21.1	14.5	26.1	61.9	31.4	4.8	32.5
ShuffleNetV2	53.0	30.2	2.1	29.3	61.6	19.5	0.3	28.6	55.5	24.8	18.3	27.3	69.0	31.7	10.3	31.8
SqueezeNet	59.7	33.4	7.3	35.1	64.6	26.0	9.5	39.8	62.2	34.5	25.6	33.6	72.0	42.0	25.6	38.4
EfficientNetB0	65.0	29.2	1.3	32.2	68.6	26.9	30.3	36.0	64.7	29.7	11.8	34.0	76.4	38.9	11.5	39.1
GhostNetV1	60.7	31.5	8.2	31.3	64.2	23.1	16.1	33.8	55.7	18.8	11.9	28.6	68.2	32.5	11.0	34.7

as the backbone. The results of different baseline models are shown in Table VI, Table VIII, and Table IX.

1) Cross-Camera Setting: Table VI shows the results of the cross-camera setting in the daytime. We can find that the results of each view are largely lower than that of the traditional setting. For example, when using ResNet-50 as the backbone, the model trained on the data of Camera 2 produces 68% accuracy on the testing set of Camera 2. However, the testing result is reduced to 50% when using the model trained on the data of Camera 1. These results indicate that the models significantly suffer from the variations caused by camera changes. The transfer performance is closely related to the angle difference between the two cameras. For instance, the transfer result of the model trained on the data of Camera 1 is successively decreased from Camera 2 to 4.[1] These results also suggest that we can leverage the data of internal cameras to bridge the gap between two cameras that have a large angle difference.

2) Cross-Vehicle Setting: We first list the results of cross-vehicle-type setting (*e.g.,* Sedan →SUV) on Table VII. It can be observed that the accuracy is decreased dramatically when changing the type of vehicle. This suggests that we should take into account the vehicle type in the model design. To be specific, the decline from Sedan to Van is more serious than to SUV. For instance, when using ResNet50 on Camera 1, the transfer result of Se→SUV is 36.2% while the result of Se→Van is 5.2%. We conjecture this is because the sedan and SUV share similar interior structures resulting in more similar data distribution. Therefore, to further investigate the influence of the interior structure of the vehicles, we design a cross-individual-vehicle setting (*e.g.,* Mazda →Lynk&Co) due to the interior structure may vary greatly even for the same type of vehicle. Results of cross-individual-vehicle are shown in Table VIII. We can observe that the cross-individual-vehicle accuracies are consistently decreased compared to that of training and testing with the data from all vehicles. This indicates that the individual vehicle changes will also deteriorate the accuracy even using the same camera and that we should consider the vehicle variations during training. One possible solution could be training robust models with domain generalization or domain adaptation methods.

3) Cross-Modality Setting: Since collecting daytime data is much easier than nighttime data, it is more suitable to study the transfer direction from daytime to nighttime. In Table IX, we report the results of the cross-modality setting from daytime to nighttime. Clearly, all the models produce very poor results when testing on nighttime data. This is due to the large data bias between the two modalities. In real-world applications, it is more dangerous when driving at nighttime. Therefore, it is essential to solve the

[1] The greater the difference between the ID numbers of the two cameras, the greater the angle difference between them.

TABLE IX

ACCURACY (%) OF CROSS-MODALITY SETTING. D AND N INDICATE DAY AND NIGHT, RESPECTIVELY. THE NUMBER FOLLOWING D AND N INDICATES THE CAMERA ID

Model	D1→ N1	D2→ N2	D3→ N3	D4→ N4
ResNet50	16.7	19.2	12.5	33.4
MobileNetV3	21.0	21.7	12.0	14.9
ShuffleNetV2	5.1	4.8	9.0	3.7
SqueezeNet	17.1	7.1	6.0	16.4
EfficientNetB0	13.0	7.9	9.9	21.3
GhostNetV1	12.8	6.3	3.7	5.0

TABLE X

ACCURACY (%) OF CROSS-MODALITY DOMAIN ADAPTATION.D AND N INDICATE DAY AND NIGHT, RESPECTIVELY. THE NUMBER FOLLOWING D AND N INDICATES THE CAMERA ID

Model	D1→ N1	D2→ N2	D3→ N3	D4→ N4
Source-only	16.7	19.2	12.5	33.4
DANN [53]	41.9	34.3	37.0	56.3
D-Coral [54]	38.7	40.0	38.3	56.2
BNM [55]	45.8	40.7	28.2	47.0
CDAN [56]	61.5	51.0	57.4	70.5
DSAN [57]	56.1	55.5	55.4	73.7
Supervised	66.3	55.7	53.3	74.1

cross-modality problem in DDC. We next give an effective solution to address this problem in the view of domain adaptation.

4) Cross-Modality Domain Adaptation: Domain adaptation is an effective way to address the problem of domain shift. Since different cameras, modalities, and vehicles can refer to domains, we can use domain adaptation to improve the performance of cross-camera, cross-modality, and cross-vehicle problems. In this paper, we choose the cross-modality as an example and evaluate 5 popular domain adaptation methods[2] on it, including DANN [53], D-Coral [54], BNM [55], CDAN [56] and DSAN [57]. When using domain adaptation methods, we additionally utilize unlabeled target data for training. Here, source and target data belong to different modalities. The source-only model is trained with the labeled source data while the supervised model is trained with labeled target data. Results reported in Table X show that domain adaptation methods can significantly improve cross-modality accuracy. Specifically, CDAN and DSAN achieve the best adaptation results. It is interesting that these two methods can produce slightly lower or even higher results than supervised models. This indicates that the knowledge of labeled daytime data can well be transferred to the nighttime data with a proper method and that the daytime data can be used to improve the performance on the nighttime modality. Therefore, in real-world applications, we can collect labeled daytime data and unlabeled nighttime data, and utilize effective domain adaptation methods to learn models that are robust to nighttime scenes. Considering the difficulty of annotating nighttime data, cross-modality domain adaptation can help us achieve a modality-robust model while saving labeling costs. The other two settings, *i.e.,* cross-camera and cross-vehicle, can also use domain adaptation techniques to achieve more robust models.

Research Findings.

- Camera location, individual vehicle, vehicle type, and data modality variations will deteriorate the accuracy.
- The transfer result of the model trained on Camera 1 is successively decreased from Camera 2 to 4.
- Domain adaptation methods can improve cross-modality accuracy.

Future Directions.

- Utilize the data of internal cameras to bridge the gap between two cameras that have a large angle difference.
- Study the domain generalization or domain adaptation methods for cross-domain settings in DDC.

V. CONCLUSION

In this paper, we introduce a new dataset, named 100-Driver, for Distracted driver classification (DDC). 100-Driver is the largest DDC dataset to date and is diverse in multiple important aspects. The significant properties of our dataset enable us to study 3 practical problems on 100-Driver, *i.e.,* cross-camera, cross-modality, and cross-vehicle settings that are largely overlooked in DDC as well as to explore the collaboration of multiple cameras for improving recognition accuracy. Extensive experiments conducted on 100-Driver reveal the new challenges and valuable insights/instructions to the DDC community. In summary, We hope our dataset can inspire the researchers to consider more challenges but realistic problems in DDC, pushing forward the development of safe driving monitoring systems.

Our main findings, insights, and limitations can be briefly concluded in the following. (a) **Findings.** We for the first time in the DDC field find that the camera location, vehicle type, and data modality changes can largely influence performance, and the cameras located at front-left and side-right (refers to Figure 4) could produce better accuracy. This finding confirms the necessity of a large-scale and diverse in terms of the camera view, vehicle, and data modality. In addition, this study finds that the combination of two views is able to improve the accuracy whereas the fusion of three or more views brings limited improvement and may greatly increase the complexity. This, therefore, indicates the benefit gained from multiple-camera fusion is not linear, that is, the more is not always the better. Most notably, it is found that the domain adaptation methods show the potential in dealing with the cross-modality problem. This suggests that the domain adaptation approach appears to be effective in cross-domain settings that will be encountered in practice. (b) **Insights/instructions**. Based on our findings, we suggest the user install the cameras at front-left and side-right locations. Furthermore, if the manufacturer wants to obtain a more accurate performance, it is efficient to combine the information from two cameras. (c) **limitations**. However, some limitations are worth noting. The effects of individual differences like age and driving age differences are not investigated in the current work. Future studies should therefore include follow-up work designed to evaluate whether the recognition results are different in various ages and driving age groups.

[2] We adopt the source code released by [58] to implement domain adaptation experiments.

Acknowledgment

The authors would like to thank the anonymous referees for their helpful suggestions. They would also like to thank all the volunteers participating in their project.

References

[1] *Global Status Report on Road Safety 2018: Summary*, World Health Org., Geneva, Switzerland, 2018.

[2] Y. Yan, S. Zhong, J. Tian, and L. Song, "Driving distraction at night: The impact of cell phone use on driving behaviors among young drivers," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 91, pp. 401–413, Nov. 2022.

[3] O. Oviedo-Trespalacios, M. M. Haque, M. King, and S. Washington, "Understanding the impacts of mobile phone distraction on driving performance: A systematic review," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 360–380, Nov. 2016.

[4] P. Choudhary, A. Gupta, and N. R. Velaga, "Perceived risk vs actual driving performance during distracted driving: A comparative analysis of phone use and other secondary distractions," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 86, pp. 296–315, Apr. 2022.

[5] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets," Virginia Tech Transp. Inst., Blacksburg, VA, USA, Tech. Rep. S2-S31-RW-3, 2016.

[6] M. A. Regan et al., "The Australian 400-car naturalistic driving study: Innovation in road safety research and policy," in *Proc. Australas. Road Saf. Res., Policing Educ. Conf.*, 2013, pp. 1–13.

[7] A. Persson, H. Jonasson, I. Fredriksson, U. Wiklund, and C. Ahlström, "Heart rate variability for classification of alert versus sleep deprived drivers in real road driving conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3316–3325, Jun. 2021.

[8] T. Katsuki, K. Zhao, and T. Yoshizumi, "Learning to estimate driver drowsiness from car acceleration sensors using weakly labeled data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3002–3006.

[9] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2368–2377, Dec. 2014.

[10] C.-Y. Chiou, W.-C. Wang, S.-C. Lu, C.-R. Huang, P.-C. Chung, and Y.-Y. Lai, "Driver monitoring using sparse representation with part-based temporal face descriptors," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 346–361, Jan. 2020.

[11] B. Cyganek and S. Gruszczyński, "Hybrid computer vision system for drivers' eye recognition and fatigue monitoring," *Neurocomputing*, vol. 126, pp. 78–94, Feb. 2014.

[12] J. Jo, S. J. Lee, K. R. Park, I.-J. Kim, and J. Kim, "Detecting driver drowsiness using feature-level fusion and user-specific classification," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1139–1152, Mar. 2014.

[13] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 660–665.

[14] S. Wang, Y. Zhang, C. Wu, F. Darvas, and W. A. Chaovalitwongse, "Online prediction of driver distraction based on brain activity patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 136–150, Feb. 2015.

[15] A. Kashevnik, I. Lashkov, and A. Gurtov, "Methodology and mobile application for driver behavior analysis and accident prevention," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2427–2436, Jun. 2020.

[16] F. Vicente, Z. Huang, X. Xiong, F. D. L. Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[17] B. Qin, J. Qian, Y. Xin, B. Liu, and Y. Dong, "Distracted driver detection based on a CNN with decreasing filter size," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6922–6933, Jul. 2022.

[18] W. Li, J. Wang, T. Ren, F. Li, J. Zhang, and Z. Wu, "Learning accurate, speedy, lightweight CNNs via instance-specific multi-teacher knowledge distillation for distracted driver posture identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17922–17935, Oct. 2022.

[19] J. Wang et al., "A survey on driver behavior analysis from in-vehicle cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10186–10209, Aug. 2022.

[20] Z. Wharton, A. Behera, Y. Liu, and N. Bessis, "Coarse temporal attention network (CTA-Net) for driver's activity recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1279–1289.

[21] Kaggle. (2016). *State Farm Distracted Driving Dataset*. [Online]. Available: https://www.kaggle.com/c/state-farm-distracted-driver-detection/data

[22] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," in *Proc. NeurIPS*, 2018, pp. 1–18.

[23] J. D. Ortega et al., "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Proc. ECCV*, 2020, pp. 387–405.

[24] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "MDAD: A multimodal and multiview in-vehicle driver action dataset," in *Proc. ICCAIP*, 2019, pp. 518–529.

[25] T. A. Dingus et al., "The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment," Dept. Transp., Washington, DC, USA, Tech. Rep. HS-810 593, 2006.

[26] R. Eenink, Y. Barnard, M. Baumann, X. Augros, and F. Utesch, "UDRIVE: The European naturalistic driving study," in *Proc. Transp. Res. Arena*, 2014, pp. 1–10.

[27] H. Abut et al., "Data collection with 'UYANIK': Too much pain; but gains are coming," in *Corpus and Signal Processing for Driver Behavior*. Springer, 2007, ch. 3.

[28] E. Q. Wu et al., "Novel nonlinear approach for real-time fatigue EEG data: An infinitely warped model of weighted permutation entropy," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2437–2448, Jun. 2020.

[29] L. Jin, Q. Niu, H. Hou, H. Xian, Y. Wang, and D. Shi, "Driver cognitive distraction detection using driving performance measures," *Discrete Dyn. Nature Soc.*, vol. 2012, pp. 1–12, Jan. 2012.

[30] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead—A large-scale driver head pose dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1–10.

[31] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons," in *Proc. 4th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Oct. 2012, pp. 149–154.

[32] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 818–830, Apr. 2014.

[33] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.

[34] C. Zhang, X. Lu, Z. Huang, S. Xia, and C. Fu, "A driver fatigue recognition algorithm based on spatio-temporal feature sequence," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–6.

[35] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3182–3190.

[36] G. M. Bhandari, A. Durge, A. Bidwai, and U. Aware, "Yawning analysis for driver drowsiness detection," *Int. J. Res. Eng. Technol.*, vol. 3, no. 2, pp. 502–505, 2014.

[37] S. Abtahi, B. Hariri, and S. Shirmohammadi, "Driver drowsiness monitoring based on yawning detection," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, May 2011, pp. 1–4.

[38] T. Hoang Ngan 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2016, pp. 46–53.

[39] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3017–3032, Aug. 2015.

[40] F. Omerustaoglu, C. O. Sakar, and G. Kar, "Distracted driver detection by combining in-vehicle and image data using deep learning," *Appl. Soft Comput.*, vol. 96, 2020, Art. no. 106657.

[41] T. Billah, S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, "Recognizing distractions for assistive driving by tracking body parts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1048–1062, Apr. 2019.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] A. Howard et al., “Searching for MobileNetV3,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Nov. 2019, pp. 1314–1324.

[44] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in *Proc. ECCV*, 2018, pp. 116–131.

[45] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size,” in *Proc. ICLR*, 2016, pp. 5987–5995.

[46] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, 2019, pp. 6105–6114.

[47] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “GhostNet: More features from cheap operations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proc. AAAI*, 2020, pp. 13001–13008.

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[51] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to reweight examples for robust deep learning,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[53] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. ICML*, 2015.

[54] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *Proc. ECCV*, 2016.

[55] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, “Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3940–3949.

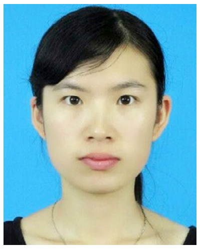
[56] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *Proc. NeurIPS*, 2018.

[57] Y. Zhu et al., “Deep subdomain adaptation network for image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1713–1722, Apr. 2021.

[58] J. Wang and W. Hou. (2021). *Deepda: Deep Domain Adaptation Toolkit*. [Online]. Available: https://github.com/jindongwang/transferlearning/tree/master/code/DeepDA

Jing Wang received the M.S. degree from the School of Physics and Technology, Wuhan University, in 2013. She is currently pursuing the Ph.D. degree with the Hefei Institute of Physical Science, Chinese Academy of Sciences, and the Graduate School of Computer Applied Technology, University of Science and Technology of China. Her current research interests include deep learning, image processing, and driver behavior analysis.

Wenjing Li received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC). She is currently a Post-Doctoral Researcher with HFIPS, Chinese Academy of Sciences. Her research interests include image and video understanding, few-shot learning, and model deployment.



Fang Li received the Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC) in 2016. She is currently an Engineer at the Hefei Institutes of Physical Science, Chinese Academy of Sciences. Her current research interests include automatic control, machine learning, and blockchain technology.

Jun Zhang received the Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC) in 2020. He is currently a Deputy Researcher at the Hefei Institutes of Physical Science, Chinese Academy of Sciences. His current research interests include the Internet of Things, machine learning, and pattern recognition.

Zhongcheng Wu received the Ph.D. degree from the Institute of Plasma Physics, Chinese Academy of Sciences, in 2001. From 2001 to 2004, he was a Post-Doctoral Researcher with the University of Science and Technology of China (USTC). He is currently a Professor with the Hefei Institutes of Physical Science, Chinese Academy of Sciences, and a Ph.D. Supervisor of both USTC and the Chinese Academy of Sciences. His research interests include sensor technology and human–computer interaction.

Zhun Zhong received the Ph.D. degree from the Department of Artificial Intelligence, Xiamen University, China, in 2019. He was a Joint Ph.D. Student with the University of Technology Sydney, Australia. He was a Post-Doctoral Researcher with the University of Trento, Italy, where he is currently an Assistant Professor. His research interests include person re-identification, novel class discovery, data augmentation, and domain adaptation.

Nicu Sebe (Senior Member, IEEE) is currently a Professor with the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He is a fellow of IAPR. He was the General Co-Chair of the IEEE FG 2008 and ACM Multimedia 2013 and the Program Chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017, and ICPR 2020. He is the General Chair of ACM Multimedia 2022 and the Program Chair of ECCV 2024.