Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles

Manuel Martin*[1] Alina Roitberg*[2] Monica Haurilet[2] Matthias Horne[1]
Simon Reiß[2] Michael Voit[1] Rainer Stiefelhagen[2]

[1]Fraunhofer IOSB, Karlsruhe [2] Karlsruhe Institute of Technology (KIT)
* equal contribution, alphabetical order

www.driveandact.com

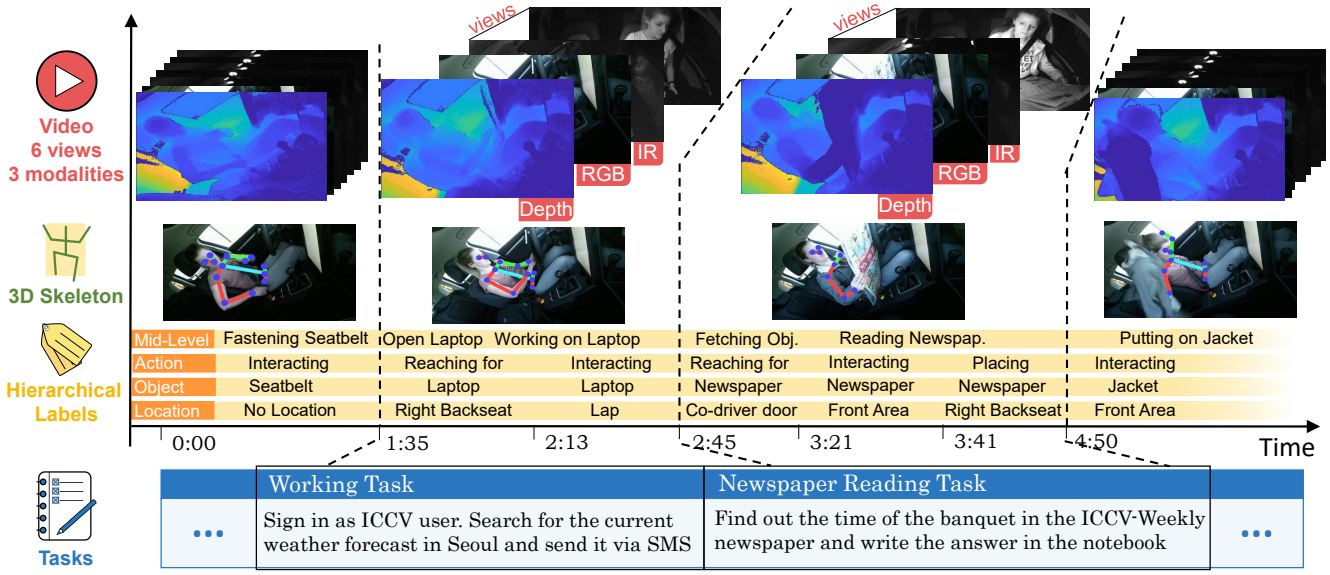

Figure 1: Overview of the Drive&Act dataset for driver behavior recognition. The dataset includes 3D skeletons in addition to frame-wise hierarchical labels of 9.6 Million frames captured by 6 different views and 3 modalities (RGB, IR and depth).

Abstract

We introduce the novel domain-specific Drive&Act benchmark for fine-grained categorization of driver behavior. Our dataset features twelve hours and over 9.6 million frames of people engaged in distractive activities during both, manual and automated driving. We capture color, infrared, depth and 3D body pose information from six views and densely label the videos with a hierarchical annotation scheme, resulting in 83 categories. The key challenges of our dataset are: (1) recognition of fine-grained *behavior inside the vehicle cabin; (2)* multi-modal *activity recognition, focusing on diverse data streams; and (3) a* cross-view *recognition benchmark, where a model handles data from an unfamiliar domain, as sensor type and placement in the cabin can change between vehicles. Finally, we provide challenging benchmarks by adopting prominent methods for video- and body pose-based action recognition.*

1. Introduction

While the rise of automation encourages distractive behavior of the driver, most of the computer vision research has been focused on understanding the situation *outside* the vehicle [13, 39, 52]. At the same time, looking at the human *inside* the cabin has strong potential to improve human-vehicle communication, dynamic driving adaptation and safety. The majority of traffic accidents involve secondary activities behind the steering wheel, an estimated 36% of such crashes could be avoided if no distraction occurred [10]. While future drivers will be gradually relieved from actively steering the car, the transition to the level of complete automation is a long-lasting process [1]. Over-reliance on automation might lead to catastrophic consequences, and, for a long time, the driver will need to intervene in case of uncertainty [1, 38, 31]. Besides identifying driver distraction for safety reasons, activity recognition

may increase comfort e.g. by adjusting the driving style if the person is drinking coffee or turning on the light, when reading a book.

Driver behavior recognition is closely linked to the broader field of action recognition, where the performance numbers have rapidly increased due to the rise of deep learning [7, 44, 46]. Such models are data-hungry and are often evaluated on large, color-based datasets with a carefully selected set of highly discriminative actions, usually originated from Youtube [7, 25]. Presumably due to the insufficient datasets for training such models, the research on driver activity understanding is far behind. Existing works are often conducted on private datasets [35, 50] and are limited to the classification of very few low-level actions (e.g. whether the person is holding the steering wheel, or switching gear [35]). None of the existing benchmarks cover higher level activities (e.g. changing clothes), especially in context of highly automated driving.

We aim to facilitate research of activity recognition under realistic driving conditions, such as low illumination and limited body visibility and present the novel *Drive&Act* dataset. *Drive&Act* offers a variety of potential challenges in connection with practical applications of activity recognition models and is the first publicly available dataset, which combines the following properties:

- **Driver secondary activities** in context of both autonomous and manual driving (83 classes in total).
- **Multi-modality**: color-, depth-, infrared- and body pose data, as conventional RGB-based action recognition datasets disregard the case of low illumination.
- **Multi-view**: six synchronized camera views cover the vehicle cabin to deal with limited body visibility.
- **Hierarchical activity labels** on three levels of abstraction and complexity, including context annotations.
- **Fine-grained** distinction between individual classes (e.g. *opening bottle* and *closing bottle*) and **high diversity** of action duration and complexity, which poses an additional challenge for action recognition approaches (e.g. *opening door from inside* often takes less than a second while *reading a magazine* might last for minutes).

In addition to autonomous driving applications, our dataset fills the lack of a large multi-modal benchmark for concise recognition on multiple levels of abstraction. An extensive evaluation of state-of-the-art approaches for video- and body pose based action recognition demonstrates the difficulty of our benchmark, highlighting the need for further extensive action recognition research.

2. Related Work

Conventional and Driver Action Recognition Conventional video-based action recognition architectures usually derive from image-based models, where the core classification is applied on video frames and extended to the temporal dimension [7, 44, 34, 20, 15]. There are different strategies for handling the additional dimension: classifying image frames with conventional 2D CNNs and then averaging the results of all frames [44], placing a recurrent neural network on top of the CNN [34, 11] or learning spatio-temporal features through 3D convolution filters [20, 46, 7]. In comparison, deep learning-based methods for driver behavior analysis often use a similar structure, while also keeping in mind other challenges that are encountered in a realistic driving scenario e.g. changing illumination conditions. Even though some of these models make use of color cameras [49, 50, 33, 12], various methods opt for illumination-invariant sensors like IR cameras [50, 30], depth sensors [8, 27, 48] or multi-modal fusion of different sensor types [32, 8, 27].

Another strategy for decoupling the varying illumination conditions is using a mid-level representation e.g. 3D skeletons characterizing the body pose of the driver. Due to the complex structure of skeletons, popular ways for flattening their representation include recurrent neural networks [32, 28] and graph networks [51]. These approaches often exploit *both structural and temporal dynamics* of body pose sequences, by making use of joint hierarchies [43], kinematic models [47], spatio-temporal joint maps [53, 29] and multiple streams [32].

Related Datasets Due to the increasing popularity of action recognition in the computer vision community, a wide range of datasets were proposed for various domains: e.g. cooking-related tasks [9, 24, 41], sports [22, 40], robotics [42, 21] or more general videos from Youtube [45, 3, 26]. In comparison, Drive&Act tackles different types of challenges in the in-car setting, where we experience scarcity of training data and difficulties when using only RGB due to the dependence on ambient light. Thus, next we focus more on *driving datasets* for action recognition and only compare our benchmark to two popular datasets for conventional action recognition [43, 7].

In Table 1, we show the specifications of Drive&Act compared to two prominent datasets for action recognition: Kinetics [7] and the multi-modal NTU [43], and, to six driving related datasets [36, 35, 50, 2]. The Kinetics Human Action Video dataset is a large-scale benchmark including 400 action classes collected from Youtube videos i.e. *RGB videos without synchronized multi-view cameras*. In comparison, the NTU RGB+D dataset [43] analyzes multiple views of the scene by providing images from three different positions captured by Kinect cameras in a laboratory. All our presented datasets for car-related action recognition include color images in manual driving mode, of which HEH [35] includes depth and D.P. [50] even has IR data. We see that most of these datasets contain only few images

	SoA conven. AR	Multi-mod. AR	Driver Activity Recognition Datsets						
	Kinetics [7]	NTU [43]	HEH [36]	Ohn et al. [35]	Brain4Cars [19]	D.P.-Night [50]	D.P.-Real [50]	AUC-D.D. [2]	**Drive&Act**
Year	2017	2016	2014	2014	2015	2016	2016	2017/18	2019
Publicly available	✓	✓	✓	–	✓	–	–	✓	✓
Manual driving	–	–	✓	✓	✓	✓	✓	✓	✓
Autonomous driving	–	–	–	–	–	–	–	–	✓
RGB/Grayscale	✓	✓	✓	✓	✓	✓	✓	✓	✓
Depth	–	✓	✓	N/A[b]	–	–	–	–	✓
NIR	–	✓	–	–	–	✓	–	–	✓
Skeleton	–	✓	–	–	–	–	–	–	✓
Video	✓	✓	✓	N/A[b]	✓	✓	✓	N/A[b]	✓
N° images	>76M	4M	N/A[b]	11K	2M	29K	18K	17K	> 9.6M
N° synch. views	1	3	1	2	2	1	1	1	6
Resolution	N/A[c]	1920×1080[a]	680×480	N/A[b]	1920×1088	640×480	640×480	1920×1080	1280×1024[d]
N° subjects	N/A[b]	40	8	4	10	20	5	31	15
Female / male	N/A[b]	N/A[b]	1 / 7	1 / 3	N/A[b]	10 / 10	N/A[b]	9 / 22	4 / 11
N° Classes	400	60	19	3	5	4	4	10	83
Multi-level annot.	–	–	–	–	–	–	–	–	✓
N° Levels	1	1	1	1	1	1	1	1	3
Continuous labels	–	–	–	N/A[b]	–	✓	✓	N/A[b]	✓
Object annot.	✓	–	–	–	–	–	–	–	✓

[a] RGB resolution, IR/Depth resolution is 512×424
[b] information not provided by the authors
[c] variable resolution
[d] NIR-camera resolution

Table 1: Comparison of driving and non-driving related datasets for action recognition. In this table, we depict the characteristics of the recording modalities, the content of the dataset and the properties of the provided reference labels.

(under 30K) with the exception of Brain4Cars [19] that includes 2 Million frames but addresses a different task of maneuver prediction (e.g. whether the driver will turn left or right in the next seconds). Furthermore, previous datasets only analyzed human behavior in manual driving mode and did not consider activities in an autonomous driving context.

In comparison to these datasets, Drive&Act includes over 9.6 *million frames*, over four times more than any other previously published dataset for driver action recognition (AR). Moreover, we annotated our dataset with fine-grained activities of 83 *classes* in total (i.e. 62 more activities than previous driver AR datasets). Our dataset is comprised of *twelve hours of video* captured by *multi-modal synchronized cameras* placed in *six different positions*. With the unique characteristics of Drive&Act (e.g. high variety of data streams, hierarchical, fine-grained annotations), we aim to push the field of driver behavior analysis further, additionally opening new challenges for general activity recognition.

3. The Drive&Act Dataset

To tackle the lack of domain-specific action recognition benchmarks, we collect and publicly release the *Drive&Act* dataset, featuring twelve hours of drivers engaged in secondary tasks while driving in manual and automated mode.

3.1. Data Collection

Even with state-of-the-art prototype vehicles for automated driving, initiating distracting driver behavior in street traffic or on a test track is not safe. The driver is required to monitor the vehicle and would otherwise put himself and surrounding pedestrians in danger. We therefore collect our dataset in a static driving simulator. The vehicle surroundings are simulated and projected on multiple screens around a converted Audi A3 with the SILAB simulation software[1]. Both manual, automated driving and take-overs can be induced in our setup. More information about the simulator setup is provided in the supplemental material.

To encourage diverse and proactive behavior, in each session, the driver was instructed to complete twelve different tasks (two instruction examples are illustrated in Figure 1). The first task comprises entering the car, making adjustments, beginning to drive manually and switching to the autonomous mode after several minutes. All following instructions (e.g. look up the current weather forecast with the laptop and report it via SMS), were given in random order on a mounted tablet. While most of the tasks are completed while driving autonomously, in every session, four *unexpected take over requests* are triggered. As a result, the journey is continued manually for at least one minute. While the sequence of coarse tasks was explicitly given, the exact way of their execution (i.e. the fine-grained activities) was left to the subject.

Fifteen people, four female and eleven male, participated in the data collection. To facilitate diversity, we selected participants of different body height and weight, as well as, different driving styles and familiarity with assistance systems and automation modes. All participants were recorded twice, resulting in 30 driving sessions with an average duration of 24 minutes. Most participants took less time during the second session, as they were familiar with the tasks, re-

[1] WIVW SILAB: https://wivw.de/en/silab

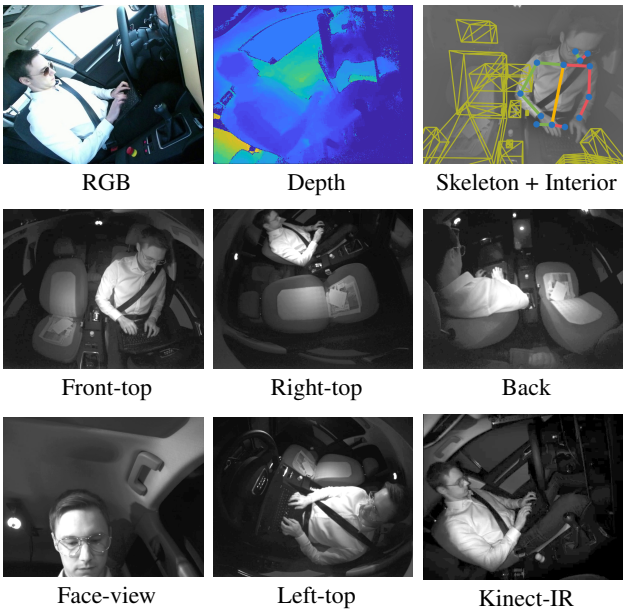

Figure 2: Example images of the *working on laptop* activity for different views and modalities.

sulting in overall different behavior and more variety in our dataset.

3.2. Recorded Data Streams

In the following, we describe the recorded data streams, covering a variety of information types, including raw video data from multiple views and modalities, 3D body- and head pose, and features capturing interactions with the car interior, which have been successfully applied for driver monitoring in the past [32].

Sensor Setup and Video Streams Two types of statically positioned cameras cover the vehicle cabin: (1) five near-infrared cameras[2] (NIR) (Resolution 1280×1024 pixel at 30 Hz) and; (2) a Microsoft Kinect for XBox One, which is used to acquire color (950×540 pixel at 15 Hz), infrared (512×424 at 30 Hz) and depth data (512×424 at 30 Hz) (Figure 2). Sensor interfaces were calibrated and synchronized with global timestamps using ROS[3]. Our setup is specifically designed for realistic driving conditions, such as low illumination. We aim to disentangle activity recognition models from conventional color input and therefore favor lightweight near-infrared cameras, which are also effective at night. Still, we acquire and release data with the Kinect sensor, which is less practical in terms of size but is very popular in the research community.

3D Body Pose To determine the 3D upper body skeleton with 13 joints, we make use of *OpenPose* [6], which is, at the time of writing, a popular choice for 2D body pose estimation. We obtain 3D poses via triangulation of 2D poses from 3 frontal views (right-top, front-top, left-top). Additional post-processing is applied to fill missing joints using interpolation of neighboring frames.

3D Head Pose To obtain the 3D head pose of the driver, we employ the popular *OpenFace* [4] neural architecture. As this model has difficulties with large head rotations, we determine the head pose on all views except for the back camera. For each frame only a subset of all cameras predict the head rotation successfully. From these candidates we pick the result of the camera with the most frontal view and transform it to world coordinates.

Interior Model We also provide car-interior features based on 3D primitives that depict interaction of the driver with his surroundings. This representation comprises location information of different storage spaces present in the car (e.g. seats or footwell) and car controls (e.g. the steering wheel, seatbelt and gear stick), which have been successfully applied for driver observation in the past [32].

Activity Classes The recorded video frames were labeled manually by a human annotator on three levels of abstraction, resulting in 83 action classes in total. We describe our hierarchical annotation scheme in detail in Section 4. It targets high-level scenarios, fine-grained activities, which retain a semantic meaning, and low-level atomic action units, which represent environment and object interactions.

3.3. Data Splits

Since we specifically aim to rate generalization to new drivers, we evaluate the models exclusively on people previously *unseen* by the classifier. We randomly divide our dataset into three splits based on the identity of the person behind the steering wheel. For each split, we use the data of ten subjects for training, of two subjects for validation, and of three drivers for testing (i.e. 20, 4 and 6 driving sessions, respectively). Since the annotated actions vary in their duration, we divide each action segment in chunks of 3s or less and use them as samples in our benchmark. We provide evaluation scripts to facilitate comparable results.

4. Hierarchical Vocabulary of Driver Actions

To adequately represent real driving situations, we conducted a thorough literature review on secondary tasks during manual driving using three types of sources: (1) driver interviews, (2) police reviews of accidents, as well as, (3) naturalistic car studies [5, 17, 23, 14]. Key factors for the choice of the in-cabin scenarios have been the *frequency* of activity engagement while driving and action *impact* on drivers attention (e.g. via increased accident odds). Furthermore, we asked five experts from the car manufacturing industry and research experts for human-vehicle interaction

[2] Camera specification: en.ids-imaging.com/store/ui-3241le.html
[3] www.ros.org

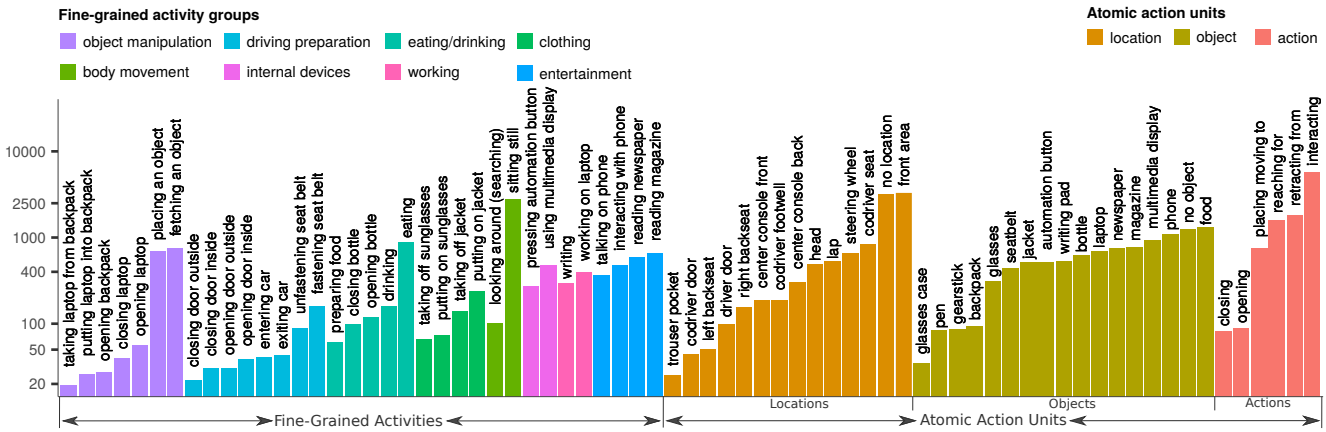

Figure 3: Sample frequency of fine-grained activities (left) and atomic actions (right) by class (logarithmic scale). A sample corresponds to a 3s snippet with the assigned label. Colors denote the activity group (e.g. food-related activities).

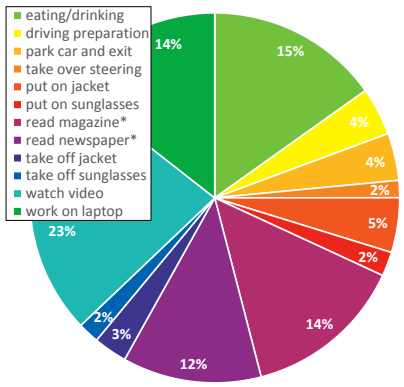

Figure 4: Distribution of the scenarios/tasks in our dataset. *these tasks consist of both finding information about a previously asked question by reading a newspaper/magazine and of writing the answer into a notebook.

to rate individual activities in terms of their usefulness for future applications.

The results indicate high interest in classes such as *talking on a mobile phone*, *working on a laptop*, *searching for something* and recognition of basic body movements (e.g. *reaching for something on the floor*), while actions such as *smoking cigarette* were rated as less useful. Certain categories, such as *sleeping*, were omitted due to technical feasibility. Following the literature review and the expert survey, we define a vocabulary of relevant driver activities from eight areas: *eating and drinking*, *clothing and accessories*, *working*, *entertainment*, *entering/exiting and car adjustment*, *body movement*, *object manipulation* and *using vehicle-internal devices*. Our final vocabulary contains 83 activity labels on three levels of granularity, building a complexity- and duration-based hierarchy of three levels.

4.1. Scenarios / Tasks

The twelve tasks our subjects had to complete in each session (Section 3.1) shape the *first level* of our hierarchy and are either scenarios typical during manual driving (e.g. *eating and drinking*) or highly distracting situations which are expected to become common with increasing automation (e.g. *using a laptop*). Figure 4 illustrates the frame-wise frequency analysis of the *scenarios* revealing that our subjects spend most of the time (23%) in the entertainment task (i.e. watching a video), and the shortest time driving manually after a take over request. The *take over* scenario is special, because the subject was unexpectedly asked to interrupt what he was doing to take over and switch to manual driving. Analyzing the reaction to such an event (e.g. in relation to prior activities or persons' age), is a potential safety-relevant research direction.

4.2. Fine-grained Activities

The *second level* represents *fine-grained activities*, breaking down the *scenarios / tasks* into 34 concise categories. In contrast to the upcoming third level of *atomic action units*, the second level classes preserve a clear semantic meaning. These fine-grained activities alternate freely during a scenario i.e. the driver is *not* told *how* to execute the task in detail. Of course, there is a strong causal link between different degrees of abstraction, as composite behaviors often comprise multiple simpler actions.

A key challenge for recognition at this level is the concise nature of the classes, as we differentiate between *closing bottle* and *opening bottle* or between *eating* and *preparing food*. We argue, that such detailed discrimination is important for applications, as the coarse components of the scene (i.e. the vehicle cabin or the loose body position) often remain similar and the relevant class-differences occur at a smaller scale than in traditional action recognition

benchmarks. As a consequence of such detailed annotation the frequency of individual classes is varying, as seen in Figure 3, which presents an analysis of the class distribution. On average, our dataset features 303 samples per class, with *taking laptop from backpack* being the least represented (19 samples) and *sitting still* being the most frequent category (2797 samples). While we refer to the three second chunks as our samples (Section 3.3), the duration of complete segments varies greatly depending on the activity (Figure 5).

4.3. Atomic Action Units

The annotations of *atomic action units* portray the lowest degree of abstraction and are basic driver interactions with the environment. The action units are detached from long-term semantic meaning and can be viewed as building blocks for complex activities of the previous levels. We define an atomic action unit as a triplet of *action*, *object* and *location*. We cover 5 types of actions (e.g. *reaching for*), 17 object classes (e.g. *writing pad*) and 14 location annotations (e.g. *co-driver footwell*), with their distribution summarized in Figure 3. Overall, 372 possible combinations of action, object and location were captured in our dataset.

4.4. Additional Annotations

We further provide dense annotations of the driving context, indicating whether the driver is in the automated driving mode or steering with the left, right or both hands. We also include the timestamps of the take over requests and simulator-internal signals e.g. the steering wheel angle.

5. Activity Recognition Models in Context of Autonomous Driving

To better understand the performance of state-of-the-art algorithms on our dataset, we benchmark a variety of approaches and their combinations. We categorize these algorithms in two groups: (1) methods based on *body pose* and *3D features*, and (2) end-to-end methods based on *Convolutional Neural Networks (CNNs)*. While CNN-based models are often the front-runners on conventional action recognition datasets, they process very high dimensional input and are far more sensitive to the amount of training data and domain shifts, such as camera view changes. In the following, we describe both groups of methods in detail.

5.1. End-To-End Models

In image-based action recognition, the model operates directly on the video data i.e. intermediary representations are not explicitly defined, but learned via CNNs. Next, we describe three prominent CNN-based architectures for action recognition, which we adopt to our task.

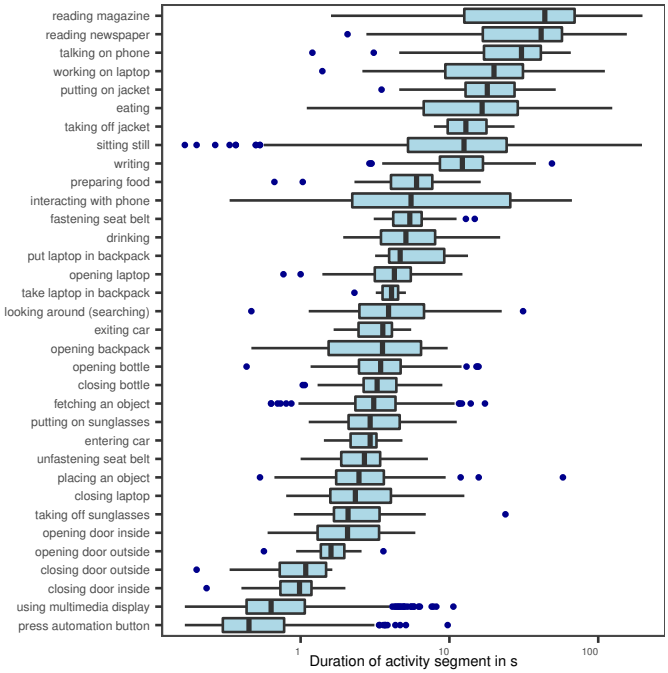

Figure 5: Duration statistics of the second-level *fine-grained activities* as boxplot (logarithmic scale). Some class names were slightly shortened due to space reasons.

C3D The C3D model [46] is the first widely-used CNN leveraging 3D convolutions for action recognition. C3D consists of 8 convolutional layers ($3 \times 3 \times 3$ kernels) and 5 pooling layers followed by two fully-connected layers.

Inflated 3D ConvNet The state-of-the-art in action recognition is currently held by the Inflated 3D architecture (I3D) proposed by Carreira et al. [7]. The architecture builds upon the Inception-v1 network [18] by extending the 2D filters with an additional temporal dimension.

P3D ResNet Unlike previous models, the P3D ResNet [37] architecture simulates 3D convolutions using $3 \times 3 \times 3$ kernels by combining a filter on the spatial domain (i.e. $3 \times 3 \times 1$) with one in the temporal dimension (i.e. $1 \times 1 \times 3$). Furthermore, P3D ResNet leverages residual connections due to their effectiveness in the field of action recognition.

5.2. Body Pose and Car-Interior Architecture

The 3D body pose is able to provide informative cues about the current activity of the driver, while still keeping human interpretability, in contrast to the mid-level feature maps produced by CNN-based architectures. Thus, we adopt the skeleton-based approach of [47] to our task that combines a *spatial* and a *temporal stream* to jointly model the body dynamics and the skeleton spatial configurations. Each stream consists of a stacked two-layered Long

Short-Term Memory (LSTM) Unit [16] followed by a fully-connected layer with softmax activation. This architecture was already adapted by Martin et al. [32] for driver action recognition by extending the network with *car-interior information* to a *three stream architecture*. In the following, we describe the input to each of the three streams:

Temporal Stream To encode the motion dynamics of the drivers' body, in each time step we unite all 13 joints via concatenation and use the produced vector in the first stream of our architecture.

Spatial Stream The second stream encodes the spatial dependencies of the joints by providing a representation of a single joint to the recurrent network at each step. To flatten the graph-based body pose representation a *traversal scheme* is used, where the sequence of joints are selected based on adjacency relations as introduced in [28].

Car-Interior Stream Since the placement of objects in the scene can provide an important cue of the current action, we also provide a representation of the interior of the car to the model. To make use of this data we determine the distance of the hands and the head to the surface of every object provided in the interior model of our dataset. This helps the network to learn the relationship between the interior of the car and the performed action.

Combined Models Following the approach of [47] for generic action recognition we combine the temporal and spatial stream with weighted late fusion. This model is called *Two-Stream* in the following. The extended model of Martin et al. [32] adds the Car-Interior network as a third stream for driver action recognition. We call this model *Three-Stream* in the following.

6. Benchmarks and Experiment Results

In the current version of our benchmark, we focus on fine-grained *classification* of driver behavior and its extension to multi-modal and cross-view settings. Given an action segment of 3 seconds or less (in case of shorter events), our goal is to assign the correct activity label. We follow standard practice and adopt the *average per-class accuracy* by using the mean of the top-1 recognition rate for every category. Note, that the random baseline is annotation level-specific and varies between 0.31% and 16.67%. In the following we focus on the performance of our baseline models. The parametrization of all models can be found in the supplemental material.

6.1. Driver Action Recognition

We evaluate our models separately for every hierarchy level: 12 scenarios/tasks (first level), 34 fine-grained activities (second level) and atomic action units with 372 possible combinations of the {*Action, Object, Location*} triplets (third level). Because the amount of triplet combinations is very high, we also report the performance for correctly classified Action, Object and Location separately (6, 17 and 14 classes, respectively).

Type	Model	Validation	Test
Baseline	Random	2.94	2.94
Pose	Interior	45.23	40.30
	Pose	53.17	44.36
	Two-Stream [47]	53.76	45.39
	Three-Stream [32]	55.67	46.95
End-to-end	C3D [46]	49.54	43.41
	P3D ResNet [37]	55.04	45.32
	I3D Net [7]	**69.57**	**63.64**

Table 2: Fine-grained Activities recognition on the Drive&Act validation and test set. We group our proposed models into: (1) baselines, (2) networks that only use the body pose representation and (3) CNN-based end-to-end methods that make predictions directly on the input images.

Fine-grained Activities In Table 2, we compare a multitude of published approaches for recognizing *fine-grained activities*, including three CNN-based methods and four models based on body- and interior representation. Overall, we achieve a mean per-class accuracy between 40.3% and 63.64%, compared to 2.94% of the random baseline. The Inflated 3D Model yields the best recognition rate (63.64%), while 3D body pose based approaches clearly benefit from combining information streams, with the Three-Stream approach being most effective in this group (46.95%). Even though we include fewer classes than the multi-modal NTU RGB+D dataset, we see that the Two-Stream model of Wang et al. [47] shows lower performance on Drive&Act, highlighting the difficulty of our benchmark.

Atomic Action Units Classification Table 3 reports the results of the *atomic action units* classification, where we show the performance of each value in the {Action, Object, Location} triplet separately, as well as, the overall accuracy of the triplet values combined. Not surprisingly, the body pose-based approaches are the front-runners for the *location* classification (56.5%), as the Three-Stream method leverages information about the interior. Moreover, the end-to-

Model	**Action**		**Object**		**Location**		**All**	
	val	test	val	test	val	test	val	test
Random	16.67	16.67	5.88	5.88	7.14	7.14	0.39	0.31
Pose	57.62	47.74	51.45	41.72	53.31	52.64	9.18	7.07
Interior	54.23	49.03	49.90	40.73	53.76	53.33	8.76	6.85
Two-Stream	57.86	48.83	52.72	42.79	53.99	54.73	10.31	7.11
Three-Stream	59.29	50.65	55.59	45.25	**59.54**	**56.5**	11.57	8.09
I3D Net	**62.81**	**56.07**	**61.81**	**56.15**	47.70	51.12	**15.56**	**12.12**

Table 3: Recognition of Atomic Action Units defined as {*Action, Object, Location*} triplets.

Type	Model	Validation	Test
Baseline	Random	8.33	8.33
Pose	Interior	35.76	29.75
	Pose	37.18	32.96
	Two-Stream	39.37	34.81
	Three-Stream	41.70	**35.45**
End-to-end	I3D Net	**44.66**	31.80

Table 4: Recognition of the coarse scenarios/tasks.

end methods often employ pooling, causing a loss of exact location information. Since the body pose based approaches do not use a visual representation of the surrounding objects, the CNN-based methods show better results for *object* (56.15%) and *action* classification (56.07%).

Scenarios/Task Recognition Table 4 shows the results of the task classification. The body-pose based approach shows better results, while the overall recognition rate is lower than in other levels. Due to the high abstraction level, we presume that the recognition would strongly benefit from a time window longer than the current 3s segments.

6.2. Multi-View and -Modal Action Recognition

In Table 5, we report the performance of the CNN-based I3D approach for the individual views and modalities and their combinations through averaging of the Softmax output scores. As expected, the recognition success correlates with the general scene visibility (see the regions covered by the cameras in Figure 2). For example, the face view setting achieves the lowest performance (42.98%) as mostly only the face of the driver is visible in this view. In comparison, the front top camera is a frontal view of the driver capturing the face, the body and close objects. While the best single-view results are achieved using the Kinect IR data (64.98%), late fusion of multiple inputs consistently improves the recognition (69.03% using all sources).

Camera	View	Validation	Test
NIR Cameras	front top	69.57	63.64
	right top	65.16	60.80
	back	54.70	54.34
	face view	49.73	42.98
	left top	68.72	62.83
	combined	72.70	67.17
Kinect Color	right top	69.50	62.95
Kinect Depth		69.43	60.52
Kinect IR		72.90	64.98
Combined		73.80	68.51
All combined (score averaging)		**74.85**	**69.03**

Table 5: Fine-grained activity level results for different views and modalities and their combination (I3D model).

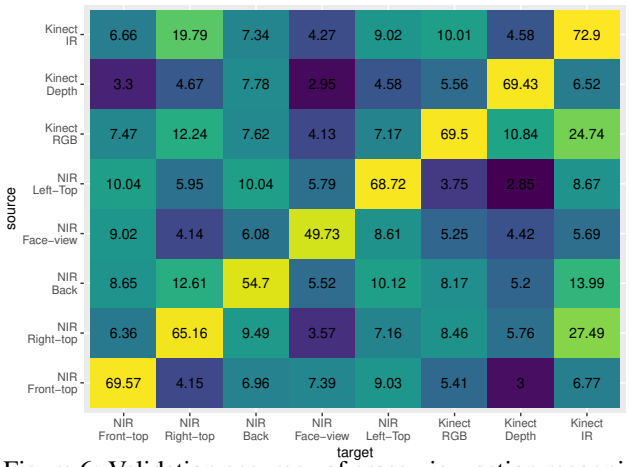

Figure 6: Validation accuracy of cross-view action recognition: the I3D model trained on data from *source* is evaluated on the *target* view. Note, that random baseline is at 2.49%.

6.3. Cross-View Action Recognition

Our next area of investigation is the cross-view and cross-modal setting, where we evaluate our best performing end-to-end method on a view not previously seen during training (results in Figure 6). Cross-view recognition is a very hard task and the performance drops significantly. Still, in most cases the models achieve better results than the random baseline. 27.49% of the fine-grained activities were correctly identified in the *Kinect IR to right top NIR* view setting and 24.74% in the cross-modal *Kinect color to Kinect IR* setting. Our results demonstrate the sensitivity of modern CNN-based action recognition models to domain shifts and highlight the need for further research of methods for handling such changes.

7. Conclusion

We present the first large-scale dataset for driver activity recognition captured in both *manual and autonomous driving mode*. The Drive&Act benchmark includes 9.6 *Million frames* captured by *six different views* and *three modalities* that were collected by five NIR- and the popular Kinect v2 cameras. The twelve hours of video are annotated by a *hierarchical annotation scheme* ranging from (1) coarse tasks the drivers had to perform and (2) fine-grained activities inside the vehicle cabin to (3) annotations of atomic action units as triplets including: the drivers' current action, the object with which the subject interacts and the object's location. We evaluate various state-of-the-art models based on both the drivers' body pose and end-to-end architectures operating on raw views. In our experiments, we highlight the difficulty of our dataset due to the concise nature of the actions and aim to facilitate further research, bringing the activity recognition models closer to real applications.

References

[1] Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles sae j 3016, 2016. 1

[2] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. *Machine Learning for Intelligent Transportation Systems Workshop in the Conference on Neural Information Processing Systems (NeuroIPS)*, 2018. 2, 3

[3] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *International Conference on Automatic Face & Gesture Recognition*, pages 59–66, 2018. 4

[5] Chad Barker. Key findings from focus group research on inside-the-vehicle distractions in new zealand. *Distracted Driving, S*, pages 213–254, 2007. 4

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 4

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3, 6, 7

[8] Céline Craye and Fakhri Karray. Driver distraction detection and recognition using rgb-d sensor. *arXiv preprint arXiv:1502.00250*, 2015. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[10] Thomas A Dingus, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016. 1

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[12] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks. In *Intelligent Vehicles Symposium (IV)*, Paris, France, June 2019. IEEE. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*. 1

[14] Craig Gordon. Driver distraction: An initial examination of the attention diverted bycontributory factor codes from crash reports and focus group research on perceived risks. 2005. 4

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*, volume 2, page 4, 2017. 2

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, 1997. 7

[17] Anja Katharina Huemer and Mark Vollrath. Ablenkung durch fahrfremde tätigkeiten" machbarkeitsstudie. 2012. 4

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 6

[19] Ashesh Jain, Hema S Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the Conference on Computer Vision*, pages 3182–3190, 2015. 3

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 2

[21] Michael Karg and Alexandra Kirsch. A human morning routine dataset. In *Proceedings of the international conference on Autonomous agents and multi-agent systems*, pages 1351–1352. International Foundation for Autonomous Agents and Multiagent Systems, 2014. 2

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[23] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006. 4

[24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the conference on computer vision and pattern recognition*, pages 780–787, 2014. 2

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2

[26] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering 12*, pages 571–582. Springer, 2013. 2

[27] Li Li, Klaudius Werber, Carlos F Calvillo, Khac Dong Dinh, Ander Guarde, and Andreas König. Multi-sensor soft-computing system for driver drowsiness detection. In *Soft computing in industrial applications*, pages 129–140. Springer, 2014. 2

[28] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016. 2, 7

[29] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, Aug. 2017. 2

[30] Tianchi Liu, Yan Yang, Guang-Bin Huang, Yong Kiang Yeo, and Zhiping Lin. Driver distraction detection using semi-supervised machine learning. *Transactions on intelligent transportation systems*, 17(4):1108–1120, 2016. 2

[31] Julina Ludwig, Manuel Martin, Matthias Horne, Michael Flad, Michael Voit, Rainer Stiefelhagen, and Sren Hohmann. Driver observation and shared vehicle control: supporting the driver on the way back into the control loop. *at - Automatisierungstechnik*, 66(2):146 – 159, 2018. 1

[32] Manuel Martin, Johannes Popp, Mathias Anneken, Michael Voit, and Rainer Stiefelhagen. Body pose and context information for driver secondary task detection. In *Intelligent Vehicles Symposium (IV)*, pages 2015–2021, 2018. 2, 4, 7

[33] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*, 2018. 2

[34] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the international conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, 2015. 2

[35] Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Mohan Manubhai Trivedi. Head, eye, and hand patterns for driver activity recognition. In *International Conference on Pattern Recognition*, pages 660–665, 2014. 2, 3

[36] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *Transactions on intelligent transportation systems*, 15(6):2368–2377, 2014. 2, 3

[37] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the International Conference on Computer Vision*, pages 5533–5541, 2017. 6, 7

[38] Jonas Radlmayr, Christian Gold, Lutz Lorenz, Mehdi Farid, and Klaus Bengler. How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 2063–2067, 2014. 1

[39] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1

[40] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *Conference on Computer Vision and Pattern Recognition*, 2008. 2

[41] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2012. 2

[42] Lukas Rybok, Simon Friedberger, Uwe D. Hanebeck, and Rainer Stiefelhagen. The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems. In *IEEE-RAS International Conference on Humanoid Robots*, 2011. 2

[43] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the International Conference on Computer Vision*, pages 4489–4497, 2015. 2, 6, 7

[47] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 7

[48] Lijie Xu and Kikuo Fujimura. Real-time driver activity recognition with random forests. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 1–8. ACM, 2014. 2

[49] Chao Yan, Frans Coenen, and Bailing Zhang. Driving posture recognition by joint application of motion history image and pyramid histogram of oriented gradients. *International journal of vehicular technology*, 2014, 2014. 2

[50] Chao Yan, Frans Coenen, and Bailing Zhang. Driving posture recognition by convolutional neural networks. *IET Computer Vision*, 10(2):103–114, 2016. 2, 3

[51] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[52] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 1

[53] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2