Activity-aware Attributes for Zero-Shot Driver Behavior Recognition

Simon Reiß* Alina Roitberg* Monica Haurilet Rainer Stiefelhagen
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany
{firstname.lastname}@kit.edu
*authors contributed equally to this work and are listed alphabetically

Abstract

In real-world environments, such as the vehicle cabin, we have to deal with novel concepts as they arise. To this end, we introduce ZS-Drive&Act *– the first zero-shot activity classification benchmark specifically aimed at recognizing previously unseen driver behaviors.* ZS-Drive&Act *is unique due to its focus on fine-grained activities and presence of activity-driven attributes, which are automatically derived from a hierarchical annotation scheme. We adopt and evaluate multiple off-the-shelf zero-shot learning methods on our benchmark, showcasing the difficulties of such models when moving to our application-specific task. We further extend the prominent method based on feature generating Wasserstein GANs with a fusion strategy for linking semantic attributes and word vectors representing the behavior labels. Our experiments demonstrate the effectiveness of leveraging both semantic spaces simultaneously, improving the recognition rate by* 2.79%.

1. Introduction and related work

While deep learning methods have demonstrated impressive results in various tasks [11, 24, 32], they are especially data-hungry. Even with enough resources to annotate a large dataset, we will never be able to capture all possible categories. One way of handling novel classes on-the-fly is Zero-Shot Learning (ZSL) [34]. ZSL connects high-level semantic descriptions of the categories to a visual model trained on the known classes to infer categories missing visual training data. The semantic description can be *e.g.* a word embedding of the category names or class-specific *attributes*. Building models that generalize well to previously unseen classes is vital for applications using *activity recognition*, which range from robotics [30] and surveillance [9] to autonomous driving [17]. While zero-shot action recognition has experienced major progress over the past years [16, 27, 29, 36], existing works have focused on recognition domains such as sports [21] or YouTube clips [12, 31]. In this work, we aim to study zero shot recognition in context of driver observation and introduce ZS-Drive&Act – the first zero-shot benchmark for activity recognition inside the vehicle cabin. We propose to leverage both textual representations of behavior labels and activity attributes, which we derive from hierarchical annotations of the Drive&Act [17] dataset and demonstrate, that such fusion consistently improves the recognition rate.

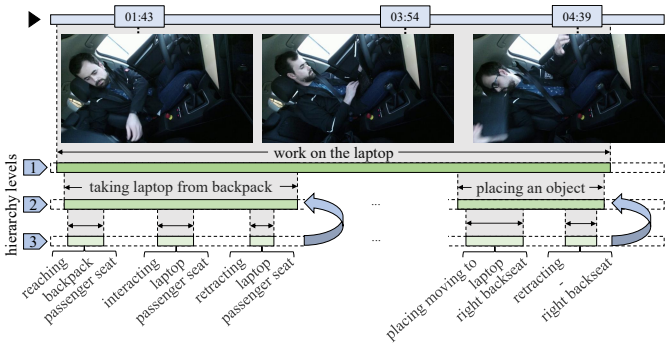

Figure 1: We leverage Drive&Act's hierarchical annotations to derive semantic attributes that model the composition of activities.

Driver behavior recognition A variety of published works have focused on driver observation, including algorithms for driver gesture recognition [23], intention prediction [8, 10], distraction detection [4, 37] and fine-grained activity recognition [17, 25, 28]. With the exception of [15], which leverages semi-supervised learning, and [28], which aims to detect novel classes, all methods study the problem in conventional closed-set supervised classification scenarios. Despite its practicality, classifying driver behaviors not present during training has not been considered yet and is therefore the main motivation of our work.

Zero-shot action recognition Early works in zero-shot action recognition [14] leverage semantic attributes that describe the actions, followed by algorithms based on textual descriptions [26], word-hierarchies [7] and word-

embeddings [36]. Recent progress in generative approaches for zero-shot image classification [5, 20, 33, 35] has sparked improvements in zero-shot action recognition, establishing a new state-of-the-art using feature generating networks [16]. Table 1 provides an overview of existing benchmarks for zero-shot activity recognition, highlighting their use of attributes and relevance for application. The Olympic Sports [21] dataset comprises sports activities, HMDB51 [12] and UCF101 [31] cover a broad corpus of more general actions, while all three datasets were collected from YouTube. These previous benchmarks require *coarse* recognition of very different activities (*i.e*. the scene context is frequently enough for the prediction). Systems in vehicles or industrial context often entail a static setting (*e.g*. car interior or a robotic cell) and the differences between the activities happen on a far more nuanced scale. *Fine-grained* recognition plays a key role in such systems and is a major challenge in our ZS-Drive&Act benchmark.

Summary and contribution This work addresses zero-shot activity recognition in context of driver observation and has three major contributions. (1) We introduce and publicly release the ZS-Drive&Act benchmark[1], the first zero-shot action recognition testbed aimed specifically at applications inside the vehicle cabin. ZS-Drive&Act is unique due to the fine-grained nature of present activities (*e.g*. *eating* and *preparing food*) and *concise activity-aware attributes*, which depict the compositional nature of behaviors and are dynamically derived from the hierarchical Drive&Act annotations (Section 2). (2) To provide a strong benchmark, we adopt and evaluate off-the-shelf zero-shot learning methods on our task covering both, attribute- and label-based algorithms (Section 3). (3) We further propose multiple enhancements for feature generating Wasserstein GANs [16], including a fusion scheme for linking the semantic spaces of attributes and word embeddings which consistently improve the recognition rate in all metrics (Section 3.2).

2. ZS-Drive&Act benchmark

To tackle the lack of zero-shot recognition datasets for automotive applications, we extend Drive&Act [17] – the largest existing dataset for conventional driver activity recognition, and introduce ZS-Drive&Act – the first zero-shot benchmark in context of driver observation. Drive&Act comprises 34 activities preformed by 15 subjects during both, autonomous and manual driving and is annotated with a hierarchical annotation scheme, where the activities are further decomposed in atomic action units.

2.1. Task description

In zero-shot driver behavior recognition, during training, a set of video instances $X^s \subseteq X$ with associated labels from a set of seen classes $Y^s \subseteq Y$ are provided. The zero-shot task entails recognizing a set of videos $X^u \subseteq X$ associated with an *unseen* set of categories $Y^u \subseteq Y$, where the zero-shot condition: $Y^s \cap Y^u = \emptyset$ holds. To bridge the gap between the seen and unseen classes, the models leverage a semantic interpretation of each of the classes: $\varphi : Y \rightarrow S$.

[1] https://github.com/Simael/zs-drive_and_act

2.2. Evaluation protocol

Evaluation setting We split the 34 available activities in Drive&Act into 14 seen classes for *training*, 10 unseen classes for *validation* and 10 unseen classes for *testing*. To mitigate the effects of particularly easy or hard splits, we use the 14-10-10 splitting and repeat it 10 times randomly. We adopt the top-1 accuracy averaged over the unseen classes as our evaluation metric (as in [34]) and report the mean and standard deviation over the ten splits.

Visual data For the videos in ZS-Drive&Act, we implement two setups: (1) we follow [17] and operate on 3 second video chunks containing some activity class and (2) we use the entire segments of activities. The latter setup enables the exploration of modeling activities with varying temporal extend as the segments span over entire activities from beginning to end. With multiple camera views and modalities available in Drive&Act, we use conventional color videos.

Pre-training By pre-training the deep learning models for conventional activity recognition on large-scale datasets such as Kinetics-600 [2], the performance of the networks increases considerably. Thus, we pre-train all networks on Kinetics-600, however, to honor the zero-shot condition we drop 56 classes related to the activities in Drive&Act.

Semantic spaces We place the Drive&Act dataset into the zero-shot setting and offer results by leveraging the *Word2Vec* [18, 19] representation of our classes. On top of this, we describe a novel scheme to directly embed the composition of activities into a semantic space.

2.3. Driver activities with attributes

The annotation hierarchy of Drive&Act (Figure 1) covers three levels: (1) *coarse tasks* – the overall task the subjects were asked to solve, (2) *fine-grained activities* – fine compositions of the tasks and, (3) the *action-units* – depicting primitive interactions with the environment. In order to derive a semantic space for the 34 *fine-grained activities* of the annotation hierarchy level (2), we leverage the *action-units* provided in level (3). The *action-units* are defined as triplets of motion-, object- and location patterns and co-occur with varying activities in the hierarchy level (2). Thus, the overlap of temporal annotations between the 34 *fine-grained activities* and the 374 different *action-units* can give insight into their characteristic composition.

	Olympic Sports	HMDB51	UCF101	ZS-Drive&Act
Clips	800	6,766	13,320	10,333
Classes	16	51	101	34
Attributes	✓	✗	✓	✓
Domain	Sports	YouTube	YouTube	Vehicle cabin

Table 1: The novel ZS-Drive&Act benchmark next to previously used zero-shot action recognition datasets.

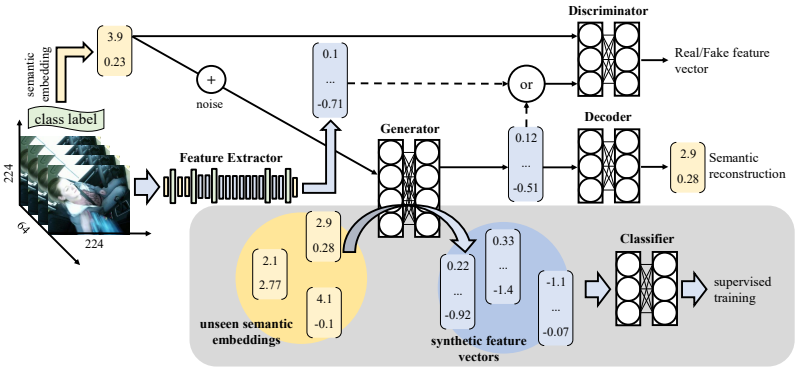

Table 2: Feature generating Wasserstein GANs re-frame ZSL into supervised recognition, by synthesizing a dataset (gray area) using semantic embeddings [16].

First, we decompose the *action-units* into their basic parts, spanning 6 motion patterns, 17 objects and 14 locations. In order to obtain attributes, we then count the frames in Drive&Act where both the activity a_i and a motion-, object- or location pattern att_j are simultaneously present. If the counts are above a threshold θ, the binary attribute is present for the current activity:

$$f_j^i = \mathbb{1}(c(a_i, att_j) > \theta) \tag{1}$$

with f^i denoting the attribute vector for activity a_i while the subscript indexes the attributes. Here, $c(a_i, att_j)$ is the number of frames in the dataset, where both a_i and att_j are annotated. Deriving the attribute vectors for Drive&Act in this fashion, 37 dimensional attribute vectors for the 34 fine-grained activities make up the semantic space, which we name *Driver Activities with Attributes* (DAwA).

3. Zero-shot driver action recognition

3.1. Zero-shot models

ConSE [22] ConSE leverages word embedding representations of the class labels and a visual model optimized for classification of the seen classes. During inference, a new representation is formed by convex combination of word embeddings of the seen classes weighted by their posterior probability. Then, the unseen activity with label closest to the previously computed vector becomes our prediction.

DAP [13] DAP leverages the presence or absence of relevant attributes within the given video. The set of attributes present in the visual input are estimated by a trained attribute recognition network. During inference, the output of the model is the best fit between the predicted attributes and the attribute vectors of the unseen classes.

DeViSE [6] The word embedding-based method DeViSE is trained to directly regress semantic vectors corresponding to the seen classes. After training on seen classes, for an input from unseen classes, a word vector is regressed and the nearest neighbor in terms of cosine-similarity is predicted.

f-WGAN [5, 16, 35] f-WGAN reframes the zero-shot action recognition problem to a supervised learning task on synthesized feature vectors (Figure 2). To this end, a conditional Wasserstein GAN [1] is trained with error-signals provided by a discriminator tasked with exposing forged feature vectors. A decoder on top of the synthesized feature vectors reconstructs the semantic embeddings to ensure that class-relevant information is retained. A synthetic dataset is set up by providing semantic embeddings of unseen classes to the generator along with noise for diverse synthetic feature vectors. Consequently, a classifier for the unseen classes is trained on this labeled generated dataset.

3.2. f-WGAN improvements

Fine-tuning We explore the effect of fine-tuning the feature extractor in f-WGANs on seen classes of the target dataset. As such, we exchange the conventional feature extractor with its fine-tuned version using visual data of seen classes.

Early Semantic Fusion The f-WGAN framework conditions its generator and discriminator on the semantic embedding $\varphi(\cdot)$. We introduce an *early semantic fusion* strategy that leverages multiple semantic spaces simultaneously. Therefore, we construct a fused semantic embedding $\varphi^*(y)$:

$$\varphi^*(y) = \varphi_1(y) \oplus \ldots \oplus \varphi_n(y) \tag{2}$$

where $\varphi_i(y)$ represents the i^{th} semantic embedding of class y and $\oplus$ denotes the concatenation operator.

Late Semantic Fusion Our second approach to fuse multiple semantic spaces trains n f-WGANs, each exploiting different semantic embeddings $\varphi_i(\cdot)$, and n corresponding classifiers c_i. The zero-shot classification result is estimated by the average over all predictions of the n classifiers.

	W2V	DAwA	#Params	3 Second Chunks		Segments	
				validation	testing	validation	testing
Random Baseline			–	10.00	10.00	10.00	10.00
CONSE	✓		12.3M	38.49 ± 10.23	28.79 ± 6.08	40.65 ± 10.72	30.01 ± 6.45
DeViSE	✓		12.6M	37.87 ± 10.59	27.69 ± 8.66	40.65 ± 11.05	28.01 ± 8.37
DAP		✓	12.3M	37.58 ± 7.08	28.66 ± 7.15	39.44 ± 6.64	27.61 ± 7.76
f-WGAN	✓		51.8M	36.64 ± 7.17	25.18 ± 4.17	41.72 ± 9.07	24.29 ± 6.66
f-WGAN + fine-tuning	✓		51.8M	37.63 ± 8.94	27.01 ± 5.27	45.28 ± 10.03	28.93 ± 7.61
f-WGAN		✓	47.6M	32.49 ± 7.45	25.04 ± 6.77	38.31 ± 8.29	26.68 ± 6.82
f-WGAN + fine-tuning		✓	47.6M	36.37 ± 8.43	26.28 ± 8.46	40.96 ± 8.08	29.96 ± 6.82
Early Semantic Fusion	✓	✓	52.5M	$\mathbf{40.28 \pm 10.39}$	$\mathbf{29.22 \pm 8.01}$	$\mathbf{46.37 \pm 10.21}$	$\mathbf{32.80 \pm 5.80}$
Late Semantic Fusion	✓	✓	99.4M	37.94 ± 9.77	28.45 ± 6.26	44.78 ± 9.13	32.55 ± 7.05

Table 3: Zero-shot driver behavior recognition results on the ZS-Drive&Act benchmark, using the 14-10-10 cross-validation split and either 3 second chunks or full activity segments. Average top-1 accuracy is reported $\pm$ the standard deviation over 10 random experiments.

Therefore, late fusion comprises an ensemble of n classifiers which are rooted in n distinct semantic embeddings.

4. Evaluation

Implementation details We train ConSE, DAP and DeViSE in an end-to-end fashion in combination with an Inflated 3D network (I3D) [3]. While ConSE comprises a standard I3D classifier with 14 output neurons and leverages Word2Vec as semantic space, DAP uses an I3D backbone with 37 neurons augmented with sigmoid activation. The sigmoid normalization of the final output layer produces values between 0 and 1, indicating the presence of the DAwA attributes. For DeViSE, we alter the final I3D-layer to have the same dimensionality as the semantic embedding (*i.e.*, in our case 300 neurons). In case of f-WGAN, we extract multiple visual features from the last I3D pooling layer averaged over the temporal dimension (the feature vector's size is 1024). For the generator, discriminator and decoder, we employ fully connected layers and showcase f-WGANs on both Word2Vec and DAwA semantic spaces. All I3D networks are initialized with pre-trained weights as described in Section 2.2. The I3D backbones of ConSE, DAP and DeViSE are fine-tuned on the seen Driver&Act classes. While we also evaluate a version of f-WGAN, where the encoder is fine-tuned on Drive&Act, the weights of encoder remain *fixed* during the generator and discriminator training. For semantic fusion approaches we fuse our proposed DAwA attributes and Word2Vec embeddings. We consider two evaluation settings: using the complete activity segments as our samples or decomposing them into three second chunks to give more weight to longer activities.

Results First, based on the results of the zero-shot models in Table 3, both DAwA and Word2Vec are suitable semantic spaces for describing the 34 activities in Drive&Act. Interestingly, the state-of-the-art f-WGAN model for zero-shot action recognition falls short in comparison to the conventional ZSL models. However, when fine-tuning the feature extractor, its performance increases up to competitive results. When incorporating DAwA and Word2Vec by means of our late fusion strategy, an increase of 2.54% over baseline models is achieved on the activity segments. Our f-WGAN enhancement with early fusion of the attribute- and word vector representation is able to boost the recognition by 4.04% (3 second chunks) and 6.12% (segments) respectively. This hints at a complementary nature of the semantic spaces as they are derived from (1) compositions of activities and (2) large text corpora. Overall, using f-WGAN together with our proposed fusion of the attribute- and word vector representation leads to the best recognition results.

5. Conclusion

Adaptive vision-based recognition infrastructure in vehicles lies vital groundwork for downstream safety related tasks such as driver monitoring. Prevalent problems when integrating deep learning into such real-world applications, are (1) annotating a large quantity of samples and (2) coping with changing environments. We address these challenges by placing the zero-shot driver behavior recognition task into the application-driven setup of Drive&Act. With our established protocol, novel video-based attributes and implementations, we showcase the relevance of this new benchmark, as the state-of-the-art zero-shot action recognition models experience major challenges. Nonetheless,

our proposed semantic fusion strategies increase the performance of the f-WGAN by 6.12%, yielding an overall performance of 32.80% on the Drive&Act test set.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017. 3

[2] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308. IEEE, 2017. 4

[4] Céline Craye and Fakhri Karray. Driver distraction detection and recognition using rgb-d sensor. *arXiv preprint arXiv:1502.00250*, 2015. 1

[5] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2018. 2, 3

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2121–2129, 2013. 3

[7] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yueting Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision (IJCV)*, 120(1):61–77, 2016. 1

[8] 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)*, pages 969–974. IEEE, 2019. 1

[9] Kyaw Kyaw Htike, Othman O Khalifa, Huda Adibah Mohd Ramli, and Mohammad AM Abushariah. Human activity recognition for video surveillance using sequences of postures. In *The Third International Conference on e-Technologies and Networks for Development (ICeND2014)*, pages 79–82. IEEE, 2014. 1

[10] 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 *International Conference on Computer Vision (ICCV)*, pages 3182–3190. IEEE, 2015. 1

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 1

[12] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *International Conference on Computer Vision (ICCV)*, pages 2556–2563. IEEE, 2011. 1, 2

[13] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(3):453–465, 2013. 3

[14] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344. IEEE, 2011. 1

[15] 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 (ITS)*, 17(4):1108–1120, 2015. 1

[16] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9985–9993. IEEE, 2019. 1, 2, 3

[17] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *International Conference on Computer Vision (ICCV)*, pages 2801–2810. IEEE, 2019. 1, 2

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119, 2013. 2

[20] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 2188–2196, 2018. 2

[21] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, pages 392–405. Springer, 2010. 1, 2

[22] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations (ICLR)*, 2013. 3

[23] 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 (ITS)*, 15(6):2368–2377, 2014. 1

[24] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 1

[25] Simon Reiß, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. Deep Classification-driven Domain Adaptation for Cross-Modal Driver Behavior Recognition. In *Intelligent Vehicles Symposium (IV)*. IEEE, 2020. 1

[26] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 46–54, 2013. 1

[27] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed Democracy: Voting-based Novelty Detection for Action Recognition. In *British Machine Vision Conference (BMVC)*, Newcastle upon Tyne, UK, September 2018. 1

[28] Alina Roitberg, Chaoxiang Ma, Monica Haurilet, and Rainer Stiefelhagen. Open Set Driver Activity Recognition. In *Intelligent Vehicles Symposium (IV)*. IEEE, 2020. 1

[29] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a fair evaluation of zero-shot action recognition using external data. In *European Conference on Computer Vision (ECCV)*. Springer, 2018. 1

[30] Alina Roitberg, Alexander Perzylo, Nikhil Somani, Manuel Giuliani, Markus Rickert, and Alois Knoll. Human activity recognition in the context of industrial human-robot interaction. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10. IEEE, 2014. 1

[31] 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. 1, 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 1

[33] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI Conference on Artificial Intelligence*, 2018. 2

[34] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1, 2

[35] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5542–5551, 2018. 2, 3

[36] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015. 1, 2

[37] Chao Yan, Frans Coenen, and Bailing Zhang. Driving posture recognition by convolutional neural networks. *IET Computer Vision*, 10(2):103–114, 2016. 1