Traffic Control Gesture Recognition for Autonomous Vehicles

Julian Wiederer[1,2,†], Arij Bouazizi[1,2,†], Ulrich Kressel[1], Vasileios Belagiannis[2]

***Abstract*— A car driver knows how to react on the gestures of the traffic officers. Clearly, this is not the case for the autonomous vehicle, unless it has road traffic control gesture recognition functionalities. In this work, we address the limitation of the existing autonomous driving datasets to provide learning data for traffic control gesture recognition. We introduce a dataset that is based on 3D body skeleton input to perform traffic control gesture classification on every time step. Our dataset consists of 250 sequences from several actors, ranging from 16 to 90 seconds per sequence. To evaluate our dataset, we propose eight sequential processing models based on deep neural networks such as recurrent networks, attention mechanism, temporal convolutional networks and graph convolutional networks. We present an extensive evaluation and analysis of all approaches for our dataset, as well as real-world quantitative evaluation. The code and dataset is publicly available[3].**

I. INTRODUCTION

Part of autonomous driving incorporates the vehicle interaction with humans. In urban traffic situations, the interaction engages pedestrians, school traffic patrols and traffic officers among others. The latter two examples are particularly interesting for the road traffic control. While a driver has learnt to recognise the traffic hand signals, it is not the same for the autonomous vehicle. Traffic control signals, i.e. hand gestures, need to be "taught" to the autonomous vehicle by means of learning databases. Understanding those gestures is essential for achieving proactive and safe autonomous driving.

On one hand, recent perception databases for autonomous driving, e.g. Cityscapes [1], ApolloScape [2] or Eurocity Persons Dataset [3], contain thousands of pedestrians, road users and cyclists, however, due to the rareness of gestures they lack scenarios with human-vehicle interaction. Road traffic controllers do not exist in this kind of databases. On the other hand, gesture recognition databases [4] include body-language, as well as human-human [5] and human-machine [6] interactions, but they lack of road traffic control gestures, such as stop or go. It becomes, thus, a necessity to create a public database for road traffic control gesture recognition.

In this work, we introduce the TCG dataset for road traffic control gesture recognition, targeted on autonomous vehicles. We define the gestures as a set of landmarks that belong to the general body pose, represented by a three-dimensional skeleton. The aim of the dataset is to classify the traffic control gestures in every time step from the sequential skeleton-based input. With the progress in human pose estimation [7], [8], [9] the body skeleton representation has become a standard input for gesture and activity recognition [10], [11], [12]. Moreover, it allows generalization to any kind of road traffic controller since it does not depend on the individual's appearance. Capturing outdoors skeleton-based traffic control gestures is not trivial though. Motion capture on public roads is forbidden due to road obstruction. To address this limitation, we work on a closed environment where we portray road intersections with multiple vehicles and the road traffic controller involved. Our recordings include all possible traffic control scenarios for road intersections with a large amount of human motion variance. Finally, our quantitative evaluations on real-world sequences show that our studio-based recordings capture the variance of the real-world. This is the first public dataset for traffic control gesture recognition to the best of our knowledge.

Alongside with the dataset, we examine a plethora of neural network approaches for gesture recognition from

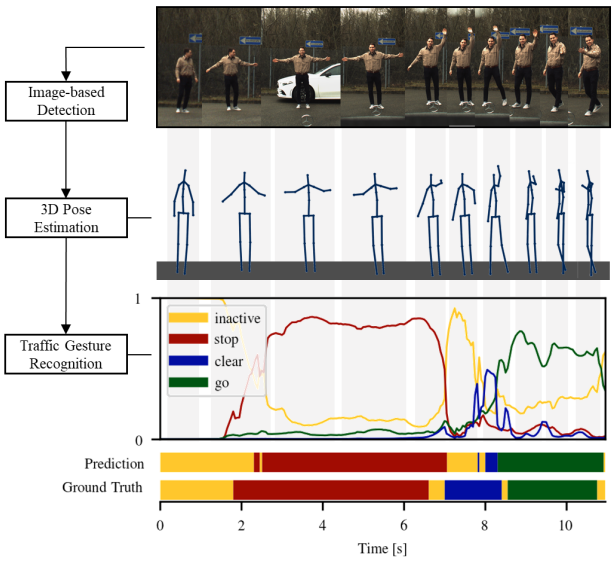


Fig. 1: Real-World Results. We demonstrate how our work functions in real-world traffic control scenarios. First, we locate the traffic controller with image-based detection and 2D body pose estimation. Second, we use 3D lifting to transform the 2D to 3D pose skeleton from a sequence of estimates. Then, the temporal gesture recognition approach predicts the gesture category based on the sequence of 3D body skeletons.

[1]Mercedes-Benz AG, Hebrhlstrae 21, 70565 Stuttgart, Germany.
[2]Universität Ulm, Albert-Einstein-Allee 41, 89081, Ulm, Germany.
† denotes equal contribution.
E-mail: *firstname.lastname@{daimler.com, uni-ulm.de}*.
[3] Project page: https://github.com/againerju/tcg_recognition

sequential data. In traffic control gesture recognition, we have a sequence to sequence problem where the gesture classification happens for each input of a 3D body skeleton. This mapping is modeled with recurrent neural networks (RNNs), including attention models, temporal convolutional networks (TCNs) and graph-based networks (GCNs). In total, we examine eight different neural networks architectures, demonstrating the advantages and limitations for each model. For that reason, we provide an extensive evaluation on our dataset and real-world sequences for cross-subject and cross-view settings, using multiple metric scores. On the real-world evaluation (see Fig. 1), we demonstrate that our dataset generalizes well outdoors, although it has been captured on a closed environment.

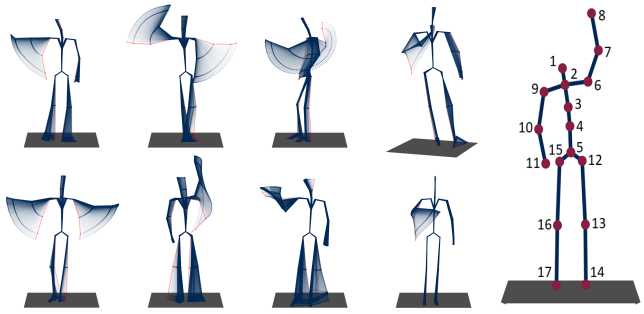
To sum up, our work makes the following contributions: 1. The first public traffic control gesture recognition dataset for autonomous vehicles. 2. An extensive evaluation of eight sequence modelling approaches, including recurrent networks, attention mechanism, TCN and GCN models. 3. An quantitative evaluation on real-world sequences to show generalization.

II. RELATED WORK

Gesture recognition for human-machine and human-human interaction is a long studied problem [13], [14]. Below, we discuss the related datasets and approaches to gesture recognition, where our focus is on human-vehicle interaction.

Human-vehicle interaction. Autonomous vehicles need to interact with humans inside the vehicle [15], [16], e.g. driver, cyclist and passengers, as well as outside the vehicle, e.g. pedestrians and police [17]. According to these studies, comprehensive understanding of the body language is important in order to react according to the human intentions. In particular, hand gestures are a common mean of interaction between the vehicle and human [18], [19]. Fortunately, the state-of-the-art on gesture recognition [20], [11], [21] allows to make easily accurate predictions. However, modeling the traffic control gestures can be challenging due to the intercultural differences [22]. For example, the traffic control hand gestures differ from country to country. In addition, road traffic control gestures are unique and they are not included in general gesture recognition datasets. In this work, we focus on the German traffic control gestures, which are also common in Europe.

Traffic control gesture recognition. Although traffic control gesture recognition becomes increasingly important in autonomous driving, the prior work on the problem is rather limited. Recently, Ma *et al*. [23] have developed a spatiotemporal convolutional neural network (CNN) to spot Chinese traffic command gestures. Similarly, a long short-term memory (LSTM) network is employed in [24] for classifying also Chinese traffic police gestures. Both approaches rely on human body skeleton input to perform the recognition. As the human body pose is in general a strong feature for activity recognition [10], [11], we also build our baselines with skeleton-based input. Compared to these prior approaches, we do not only study the problem by providing a number of algorithmic solutions, motivated by general gesture recognition, but we additionally release a public database for traffic control gesture recognition.



(a) Stop. (b) Clear. (c) Go. (d) Inactive. (e) Skeleton Representation.

Fig. 2: Our dataset is characterized by high intra-class variance. Fig. 2a to 2d show exceperts of motion sequences, in red the starting pose. In the upper image of Fig. 2d the actor is scratching her face and in the lower image someone is looking at his watch. In particular dynamic *go* gestures are challenging 2c due to their similarity to motions from the inactive class. Fig. (e) describes our 17 joint skeleton model: (1) head, (2) neck, (3) chest, (4) spine, (5) hip, (6) - (11) left shoulder, elbow, hand, thigh, knee, foot and the same for the right-hand side (12) - (17).

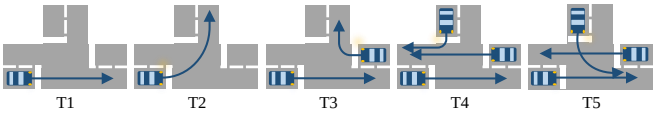
Existing gesture recognition databases. A reason for the limited research on traffic control gesture recognition is due to the lack of public data. While there are several hand gesture databases [25] for indoor scenarios [14], general gestures [26] and for specific applications such as sign language recognition [27] or egocentric gesture recognition [28]; the publicly available databases for traffic control hand gesture recognition are inexcitable. Consequently, our new public database on traffic control hand gesture supports the further research on the problem. Next, we introduce our dataset and then present the baseline algorithms for evaluation.

III. TRAFFIC CONTROL GESTURE DATASET

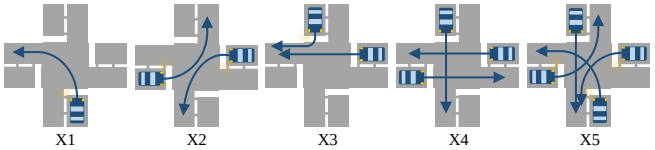
We introduce TCG, a dataset for traffic control gesture recognition, that covers all possible road traffic control variations for European road intersections. By modeling road intersections, we automatically include the non-intersection situations as well. We consider road traffic control gesture recognition as a classification task from 3D body pose skeleton input data over time. As a result, our dataset consists of 3D human body skeleton sequences represented by joint sets and the respective label per skeleton. Below, we discuss the data collection, labelling and properties, as well as the experimental setup.

A. Experimental Settings and Data Collection

We asked from 5 individuals of different body types to regulate the traffic on road intersections. We chose a T- and



(a) Five T-junction scenarios



(b) Five X-junction scenarios.

Fig. 3: Birds-eye view on the 10 scenarios of the T- and X-junction. From left to right, we observe the increasing complexity in terms of number of cars, as well as their driving intentions.

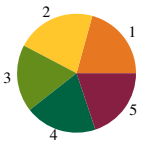
X-junction where the individual makes uses of the hands for regulation, without additional control devices like whistle or traffic paddle. We also defined 5 different scenarios for each junction, with variable number of involved vehicles. Fig. 3 shows all scenarios in bird's-eye view, while Fig. 4 presents the data distribution for all scenarios and individuals. The vehicles are specified based on their driving intention, i.e. straight, left turn or right turn, and driving order.

Since staging in real traffic situations is not permitted, we simulate the above scenarios in a closed environment, including intersection layouts, vehicles and the traffic controller. For that reason, we used colored discs to mark the streets and stopping lines. Additional colored markers were placed at the positions of the waiting vehicles to simulate the interaction partners. This helps the actors to adapt their sight according to the marker they are interacting with. In this way the setting facilitates realistic head and body orientations.

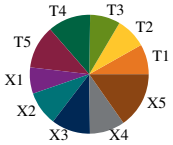
To capture the body motion of the traffic controller, each actor has been centered in the road intersection and wore an IMU[1]-based motion capture suit above the clothes. In total the suit is composed of 17 high-quality MEMS[2] inertial sensors (accelerometer, magnetometer and gyroscope) and two pressure insoles to record smooth orientation measurements in high resolution. All sensors were synchronously sampled on 100 Hz and streamed to a computer via integrated Wi-Fi transmitter. Since the computing resources are valuable and limited in an autonomous vehicle, the sampling frequencies can not be very high. For this reason, we sub-sample to 20 Hz as a reasonable frequency for autonomous vehicles. An implemented kinematic body model computes exact 3D locations and orientations of the body joints. In total, we have a skeleton model with 17 3D body joints as it is depicted in Fig. 2. Of course, the recordings would not be easily feasible outdoors. This is the advantage of the closed environment data collection.

During recordings, a lightweight script helped the actors to keep the correct order of commands, i.e. which car needs

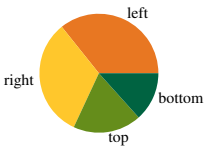
[1]Inertial Measurement Unit (IMU).
[2]Micro Electro Mechanical Systems (MEMS).



(a) Subject.



(b) Scenario.



(c) Viewpoint.

Fig. 4: Frame distribution over subjects (a), scenarios (b) and car viewpoints (c). With the scenario complexity, the sequence length is increasing.

to be stopped next and which one should proceed, while they are completely free in the duration of the commands. The script is intended as a high-level guidance rather than a detailed story line, since strong restrictions could lead to insecure and unrealistic behavior. Each scenario is repeated 5 times. In early repetitions we request the actors to perform road traffic control gestures, but after increasing the repetitions, the actors are allowed to use their own, spontaneous gestures in order to control the situation. As a starting point, all actors learn the standard European traffic control gestures, i.e. stop, go and clear. With this loose recording procedure, we achieve granularity in motion complexity, while the different actors contributed to high motion diversity.

B. Label Definition

In autonomous driving, the perception provides the environmental state, e.g. object locations or lane markings in each time step. The next action is then planned based on the history and the current state. As a result, traffic control gesture recognition should also happen continuously. To follow this principle, we build our dataset with gesture labels per time step. We reach high annotation quality with trained annotators and consequent quality-checks.

According to [22] and the German regulations, we differentiate three active gesture classes, *go*, *clear* and *stop*, as well as an *inactive* class. To increase the diversity of the *inactive* class, we actively enrich motions with daily activities, like rubbing hands, taking sunglasses on or looking at watch. Fig. 2a to 2d show examples for each class of our dataset. Additionally, the dataset provides annotation for the evaluation of transition phases, e.g. from *Stop* to *Go*. This can give insights for the decision boundaries of the gesture classifier, e.g. a gesture classifier that detects a stop gesture early in the transition phase might be a solution for autonomous driving compared to another one with larger detection latency. For the main classes, *go*, *clear* and *stop*, we sub-categorize the motion of the active hand, e.g. *left*, *right* or *both*, into *static* and *dynamic*. Fig. 5 compares the class distribution for the 5 subjects with overall 2,886 unique time intervals annotated with a major class label. The *inactive* class dominates over the classes as expected for real traffic situations. Table I provides quantitative insights of the label distribution. For most of the time, the *go* commands are

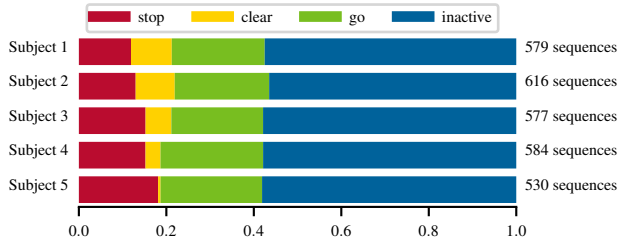


Fig. 5: The active classes constitute over 40 % of the dataset. All classes are well distributed on the 5 actors except for *clear*, which is rarely present for subject 5.

Active Classes	Stop	Clear	Go
both-hand-static	219	-	188
both-hand-dynamic	5	-	8
left-hand-static	157	32	43
left-hand-dynamic	23	-	179
right-hand-static	133	134	39
right-hand-dynamic	14	-	339
In total	551	166	796

TABLE I: In total, the TCG dataset contains 1,513 active class annotations with corresponding sub-class labels. Apparently, most of the actors where right-handed, since in more than 60 % of the one-hand gestures the right hand is used.

indicated in a dynamic way, while road traffic controllers signal stop and clear in a more static way. Dynamic stop gestures with both hands are very rare, while dynamic go with a right waving or pointing is highly present.

C. Dataset Properties

The dataset includes 250 unique 3D human body pose sequences, ranging from 16 to 90 seconds per sequence. We consider the directional property of gestures. This means that the gesture interpretation strongly depends on the viewpoint. For instance, a static stop gesture from one viewpoint will be a go gesture from another orthogonal viewpoint; or a dynamic go to the right does not mean any signal to the other participants. Therefore, the 3D body poses are transformed in the corresponding coordinate systems of the involved vehicles, i.e. the autonomous vehicle. On average, every sequence is transformed in 2.2 viewpoints, which results in 550 perspectives.

All sequences are recorded in high temporal resolution of 100 Hz and comprise 140 minutes of realistic human body motion, in total 839,350 frames. As shown in Fig. 4a, the amount of frames are evenly distributed on the 5 subjects. Apparently, with the complexity of the scenes, i.e. from *T1* to *T5* and *X1* to *X5*, sequences become longer (Fig. 4b). The pie chart over viewpoints, Fig. 4c, shows an under-representation of vehicles coming from the lower street, since it does not appear in the T-junction layout. Based on the design of the scenarios, most of the vehicles approach from the left and right.

The proposed TCG dataset can serve the community as a considerable learning base for continuous gesture recognition in the context of self-driving cars.

IV. GESTURE RECOGNITION MODELS

We define hand gesture recognition as sequence modeling, where the input sequence is the track of 3D body pose skeletons $\mathbf{x}_0, \ldots, \mathbf{x}_T$ and the output sequence is the gesture category $\mathbf{y}_0, \ldots, \mathbf{y}_T$. At each time step $t \in T$, the body skeleton $\mathbf{x}_t \in \mathcal{R}^{3\times N}$ is composed of N body joints, represented as a vector. The ground-truth gesture category $\mathbf{y}_t \in \mathcal{N}^K$ is an one-hot vector of K classes. Our goal is to learn the mapping from the input skeleton to the class category from a set of training data. Without loss of generality, we represent that mapping as:

$$\mathbf{y}_0, \ldots, \mathbf{y}_T = f(\mathbf{x}_0, \ldots, \mathbf{x}_T; \theta) \tag{1}$$

where $f : \mathcal{R}^{3\times N\times T} \rightarrow \mathcal{N}^{K\times T}$ is the mapping function. We propose to approximate the mapping function based on deep neural networks. We consider recurrent, temporal convolutional and graph convolution neural networks as three different ways to approach the problem. For all network architectures, the learning goal is to minimize the difference between the predictions and ground-truth. This can be formalized by the loss function that is given by:

$$\underset{\theta}{\operatorname{argmin}}\, L(\mathbf{y}_0, \ldots, \mathbf{y}_T, f(\mathbf{x}_0, \ldots, \mathbf{x}_T; \theta)), \tag{2}$$

that is cross-entropy for problem. Finally, the training is accomplished with back-propagation and stochastic gradient descent. Note, that we do not assume access to future time steps, i.e. $T + 1$. Next, we comment on the neural network models for each architecture type.

A. Recurrent Network Architectures

Skeleton-based action recognition approaches traditionally make use of RNNs to model the temporal dynamics. GRU-, LSTM-cells or more complex structures, such as bidirectional networks [29] are the common network architectures since vanilla RNNs do not capture long dependencies. In our evaluation, we consider all these types of RNNs for gesture recognition.

B. Attention Mechanism

Modeling long sequences can be accomplished with an attention mechanism as well. Song *et al.* [30] have shown an end-to-end spatial and temporal attention model for human action recognition. The model is trained to pay more attention on discriminative joints of the skeleton within each frame and to estimate the importance of frames in the sequence. Attention has also been used for spatiotemporal attention networks to model the evolution of dynamic hand gestures [31]. To retrieve a better semantic information, a novel model with self-attention network (SAN) was proposed by [32]. We also examine the potential of self-attention in combination with the LSTM cells.

C. Temporal Convolutional Networks

Recently, it has been shown that convolutional network architectures are on par with recurrent networks on sequence modeling [33]. At the same time, the idea of temporal

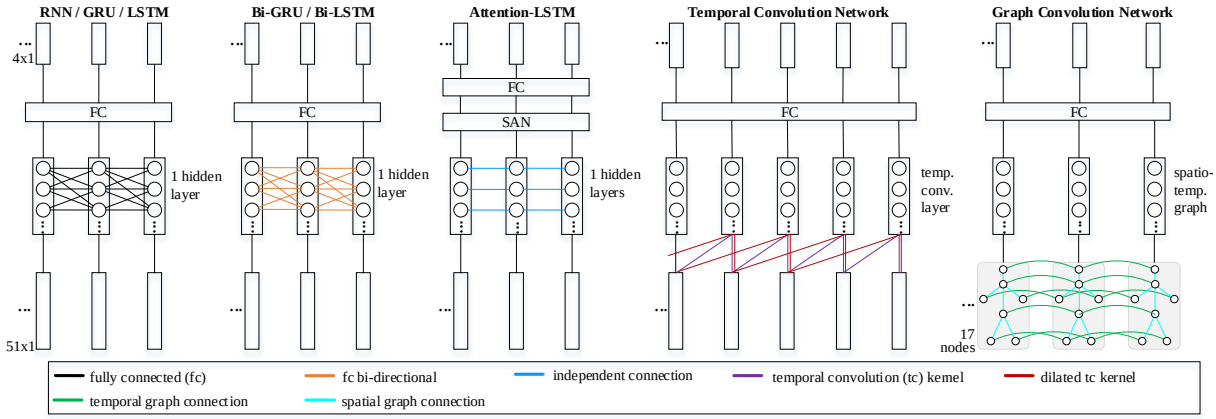


Fig. 6: Network Architectures Illustration. We show the structure of the RNN, GRU and LSTM as well as temporal and graph convolutional networks and their connectivity. The input vector for each time step is the 3D skeleton represented by 17 body joints. We refer to fully connected layer and self-attention networks as FC and SAN, respectively.

convolutions has been established for visual tasks [34], audio generation [35] and signal processing [36]. We study the effect of temporal convolutions in our problems as well. The temporal convolutions process the 3D body joints, independently, over-time.

D. Graph Convolutional Architectures

Graph neural networks are well-suited to non-structured data such as the human body, represented by a skeleton model [37]. Yan *et al.* [38] proposed a spatiotemporal graph convolutional network to perform activity recognition from skeletal data. The skeletons are composed of 2D or 3D joint positions. We rely on the same idea to perform gesture recognition. We present a graph convolutional network that processes 3D body joints to classify traffic control activities.

V. EXPERIMENTS

We evaluate the presented dataset for different sequence modelling strategies, as they have been presented in Sec. IV. The experiments include six recurrent network models, one temporal convolution network (TCN) and a spatio-temporal graph convolution network (GCN). Similar to gesture recognition approaches [39], the evaluation metrics are accuracy, as well as Jaccard index [4], F1-score and the confusion matrix. At last, we present an image-based evaluation on real-world sequences with the traffic officer and the autonomous vehicle.

A. Network Architecture Implementation

We provide the implementation details for each neural network model individually. In general, all models have been trained from scratch with grid hyper-parameter search. Moreover, the activation function is non-linear, dropout is applied everywhere with rate 0.5 and the training takes place until convergence. Class confidences are computed with a dense layer and the softmax function on top of the high-level feature representations provided by the temporal models. The optimizer is the adaptive learning rate optimization algorithm (Adam) [41], with initial learning rate 0.001, unless it is differently reported. Below, the specific configuration for each temporal model is reported.

a) Recurrent Neural Networks: We consider six types of recurrent neural networks, combined with a fully connected layer and the softmax activation function to perform gesture classification. In detail, the encoder is modeled as *vanilla-RNN*, *GRU*, *LSTM*, *bidirectional-GRU* or *bidirectional-LSTM*. Since the sequence length varies, we adopt a masking mechanism for the input 3D body skeletons as in [40] to overcome the zero-padding problem. For the bidirectional-LSTM, we adopt the architecture of [40]. For the other models, our architecture is presented in Fig. 6. In all cases, we rely on 100 cells and a single hidden layer.

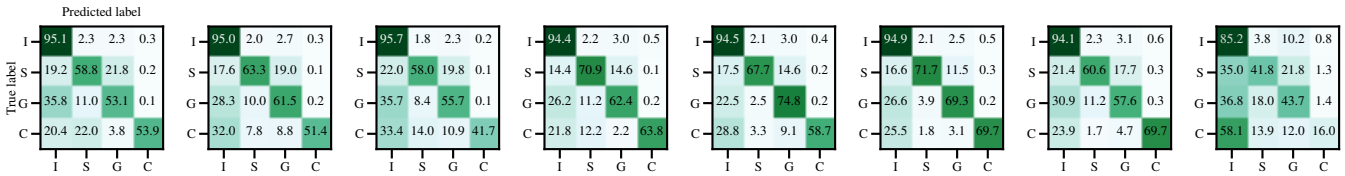
b) Attention Model: We add an attention layer on top of the *LSTM* encoder. In particular, we transform the *LSTM* to *Attention-LSTM* and make use of same architecture as before, however, empirically select 50 cells for the hidden layer and 50 attention units.

c) Temporal Convolutional Networks: We adopt [34] to implement our TCN. We build it though simpler, because it does not deal with image data. It consists of 1D convolution kernels of size 2 and 64 filters where the dilation rate goes from 2 and reaches 32, by doubling it in each layer. The parameter optimization is accomplished with Adam, with learning rate 0.001 and back-propagation. An illustration of the model is shown in Fig. 6.

d) Graph Convolutional Networks: We rely on the GCN of [38] for our problem. The body is represented as an undirected spatiotemporal graph with 17 joints and T time steps, where T=20 for making a single prediction. In the training phase we randomly sample sequences of 20 3D body skeletons of each class. During testing continuous predictions are required. Therefore we predict with a sliding window of stride 1 to obtain continuous predictions that equally compare with the other sequence modelling approaches. The initial learning rate for this model is 0.1.

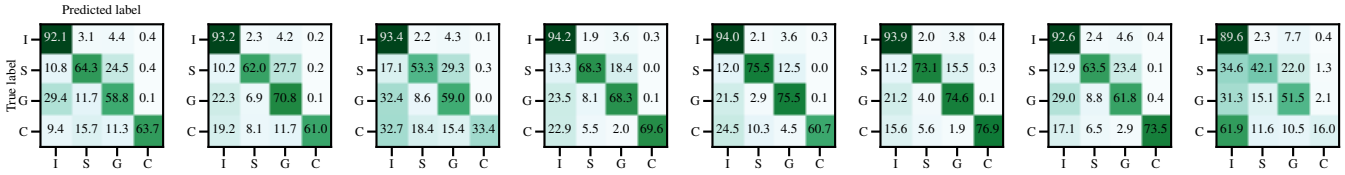
TABLE II: Results on the 4-Class Evaluation. We perform cross-subject, cross-view and real-world evaluations for all models and provide the mean and standard deviation of three runs. For all metrics, the higher score the better the result.

Methods	Cross-subject			Cross-view			Real-World		
	Accuracy	Jaccard	F1-score	Accuracy	Jaccard	F1-score	Accuracy	Jaccard	F1-score
RNN [39]	82.81 (±*2.7*)	57.40 (±*2.3*)	69.45 (±*1.4*)	80.94 (±*1.9*)	57.21 (±*2.5*)	69.98 (±*2.3*)	69.39 (±*7.2*)	39.70 (±*8.6*)	50.26 (±*10.2*)
GRU	84.44 (±*2.0*)	58.16 (±*4.2*)	70.45 (±*3.1*)	83.47 (±*1.4*)	56.25 (±*7.6*)	68.59 (±*7.4*)	71.8 (±*8.6*)	40.4 (±*10.2*)	50.67 (±*11.4*)
LSTM [39]	83.23 (±*3.6*)	56.32 (±*7.0*)	68.59 (±*6.9*)	79.58 (±*1.6*)	52.02 (±*3.2*)	64.62 (±*3.8*)	**77.88 (±*9.6*)**	**52.90 (±*15.0*)**	**62.21 (±*15.2*)**
Att-LSTM	85.67 (±*2.1*)	50.70 (±*9.9*)	61.87 (±*10.6*)	85.30 (±*1.1*)	59.87 (±*12.7*)	71.20 (±*12.3*)	72.76 (±*10.2*)	44.61 (±*15.4*)	52.50 (±*16.0*)
Bi-GRU	86.80 (±*1.6*)	57.25 (±*7.4*)	68.95 (±*6.4*)	**87.37 (±*0.3*)**	55.55 (±*2.8*)	67.68 (±*2.2*)	73.58 (±*8.1*)	43.09 (±*10.8*)	52.26 (±*12.8*)
Bi-LSTM [40]	**87.24 (±*1.8*)**	**67.00 (±*2.1*)**	**78.48 (±*1.8*)**	86.66 (±*1.2*)	**65.95 (±*4.7*)**	**77.14 (±*4.3*)**	72.28 (±*8.7*)	48.81 (±*12.5*)	61.23 (±*14.4*)
TCN [34]	83.44 (±*2.5*)	62.06 (±*2.8*)	74.23 (±*3.0*)	82.66 (±*0.7*)	63.97 (±*1.3*)	75.95 (±*0.9*)	48.70 (±*6.5*)	25.73 (±*6.4*)	35.23 (±*8.2*)
GCN [38]	65.42 (±*9.6*)	38.55 (±*13.6*)	50.73 (±*14.5*)	62.40 (±*14.2*)	35.05 (±*9.8*)	48.51 (±*11.3*)	60.64 (±*3.7*)	34.34 (±*2.9*)	48.72 (±*3.4*)



(a) Vanilla-RNN (b) GRU (c) LSTM (d) Att-LSTM (e) Bi-GRU (f) Bi-LSTM (g) TCN (h) GCN

Fig. 7: Cross-Subject Confusion Matrices on 4-Class. We abbreviate the gestures *inactive*, *stop*, *go* & *clear* as *I*, *S*, *G* & *C*.



(a) Vanilla-RNN (b) GRU (c) LSTM (d) Att-LSTM (e) Bi-GRU (f) Bi-LSTM (g) TCN (h) GCN

Fig. 8: Cross-View Confusion Matrices on 4-Class. We abbreviate the gestures *inactive*, *stop*, *go* & *clear* as *I*, *S*, *G* & *C*.

B. Cross-Subject & Cross-View Protocol

We define the cross-subject and cross-view evaluation protocol, similar to the gesture recognition approaches [39]. Note we explicitly aim to distinguish gestures dependent on a specific viewpoint, i.e. view variant recognition. In the cross-view evaluation, this means that the labels differ depending on the vehicle's viewpoint. As a result, the model is trained on all sequences of 3 viewpoints, e.g left, top and right, and evaluated on the omitted set of sequences, e.g. bottom. In the cross-subject evaluation, which is considered to be more challenging, the model is trained on 4 actors and tested on the remaining actor. The process is repeated for all combinations.

C. Real-World Experiment Description

Our ultimate goal is to make use of our dataset for real-world traffic control gesture recognition. For that reason, we captured 5 image-based sequences of real traffic control scenarios. They consist of the traffic regulator on a T-junction road intersection and the autonomous vehicle. After labeling the sequences, we perform an evaluation of the presented gesture recognition approaches.

To obtain the 3D body skeleton from the image input, first, the traffic regulator is detected and the body 2D pose is extracted with a pre-trained Mask-RNN [42] model. Since the 3D pose is necessary, we rely on the approach of Pavllo *et al.* [43] to lift the 2D body poses to 3D body pose skeletons based on a sequence of 2D poses. Second, the estimated 3D body pose skeletons are provided to the gesture recognition approach for classification. We have performed this experiment off-line for being able to follow our evaluation protocol. Our approach is outlined in Fig. 1.

D. Quantitative Evaluation

The dataset evaluation is performed for the 4-class problem, i.e. *go*, *clear*, *stop* and *inactive*. Both training and test sets come from our dataset according to the cross-subject and cross-view protocol. For the real-world evaluation, the test set is the image-based real-world sequences, as described in Sec. V-C. Since one actor of the dataset also appears in the real-world image sequences, we exclude the actor from the dataset and re-train all models. The dataset results for cross-subject, cross-view, as well as the real-world evaluation are presented in Table II. Especially in unbalanced recognition tasks, a fair metric is required to take the distribution of classes into account. For that reason we consider the Jaccard index as the most representative metric [4].

a) Cross-Subject & Cross-View: The three evaluation metrics have similar behaviour for the cross-subject and cross-view. The best performing approach is the LSTM in the bidirectional formulation for both cases as shown in Table II. Only, the accuracy of bidirectional-GRU is slightly higher than bidirectional-LSTM for the cross-view evaluation. The recurrent networks have in overall comparable performance

TABLE III: Results on the 15-Class Evaluation. We perform cross-subject, cross-view and real-world evaluations for all models and provide the mean and standard deviation of three runs. For all metrics, the higher score the better the result. We skip the results of the GCN because of poor performance.

Methods	Cross-subject			Cross-view			Real-World		
	Accuracy	Jaccard	F1-score	Accuracy	Jaccard	F1-score	Accuracy	Jaccard	F1-score
RNN [39]	78.44 (±*1.2*)	19.19 (±*3.1*)	25.33 (±*3.8*)	80.84 (±*1.0*)	24.39 (±*2.6*)	31.00 (±*2.8*)	71.28 (±*8.4*)	30.29 (±*19.8*)	33.12 (±*19.0*)
GRU	79.27 (±*1.0*)	28.59 (±*4.1*)	36.09 (±*4.7*)	81.58 (±*0.7*)	26.30 (±*0.9*)	33.74 (±*1.3*)	73.08 (±*13.2*)	30.11 (±*12.0*)	31.55 (±*12.1*)
LSTM [39]	73.26 (±*1.7*)	17.88 (±*2.9*)	22.26 (±*3.2*)	73.31 (±*0.6*)	12.98 (±*1.3*)	16.71 (±*1.7*)	**75.62 (±*8.7*)**	**40.42 (±*14.2*)**	**45.14 (±*13.1*)**
Att-LSTM	79.90 (±*1.3*)	22.92 (±*3.9*)	29.91 (±*3.9*)	83.49 (±*0.8*)	23.01 (±*3.9*)	30.73 (±*4.3*)	71.96 (±*16.2*)	32.06 (±*11.7*)	35.04 (±*11.2*)
Bi-GRU	**82.70 (±*1.1*)**	27.8 (±*4.8*)	35.9 (±*5.0*)	83.59 (±*0.9*)	25.9 (±*3.7*)	33.56 (±*4.0*)	75.17 (±*11.7*)	30.6 (±*20.5*)	33.14 (±*19.8*)
Bi-LSTM [40]	82.46 (±*0.9*)	**29.42 (±*4.7*)**	**37.77 (±*5.3*)**	**84.27 (±*1.0*)**	**27.76 (±*2.6*)**	**35.74 (±*2.7*)**	71.75 (±*12.4*)	30.80 (±*20.9*)	33.75 (±*20.1*)
TCN [34]	73.17 (±*3.8*)	11.55 (±*5.9*)	15.19 (±*7.9*)	74.84 (±*2.2*)	15.09 (±*4.9*)	19.09 (±*6.8*)	65.80 (±*8.8*)	29.75 (±*8.4*)	32.21 (±*12.4*)

other than the vanilla RNN. The temporal convolutional network has consistent results both for cross-subject and cross-view, but it is behind the recurrent models. At last, the graph convolutional network has much lower performance compared to all other models. In addition, it had difficulties to converge. We additionally provide the confusion matrices for the cross-subject (Fig. 7) and cross-view (Fig. 8) evaluation. All classifiers are able to distinct the active classes from the inactive class. Notable is the performance on *go* compared to *stop*. In most of the cases, the recognition performance is higher on the latter, which we explain with the larger amount of dynamic gestures in the *go* class.

b) Real-World: For the real-world evaluation, all metrics agree on the best performing approach as well. The LSTM model delivers the best results on the real-world sequences (see Table II), while here the bidirectional formulation does not further improve the final outcome. Next, the behavior of the temporal convolutional network is similar to the cross-subject and cross-view evaluation. In total, the real-world evaluation delivers considerable worse performance than the cross-view and cross-subject evaluation. This is expected given that the 3D body pose skeletons are algorithmically computed and thus include some sort of error.

E. Ablation Study

We consider another classification scheme of 15-class[3] problem. By moving from 4 to 15 gesture categories, our aim is to study how the static and dynamic gestures affect the classification performance. All experimental settings are the same with Sec. V-D except the loss function that is optimized for 15 classes. The results are reported in Table III. We report the results of all methods except the graph convolutional network because it has shown unstable convergence during training and thus reached poor performance.

a) Cross-Subject & Cross-View: The accuracy for cross-subject and cross-view is comparable to the 4-class problem. However, the jaccard index and F1-score show that the 15-class problem results in a descent performance reduction. This observation holds for all models. The best performing model is again the bidirectional-LSTM for both evaluations. The bidirectional-GRU accuracy is the best for the cross-subject evaluation, but the bidirectional-LSTM is in general on par with it. The other model have similar behavior by comparing the 4-class results of Table II with Table III.

b) Real-World: Unlike the cross-subject and cross-view results, the real-world performance is similar to the 4-class problem (see Table III). Considering the standard deviation, all models show great variation between runs as a result of the domain difference between the motion capture data for training and the estimated 3D poses for testing. The clear observation is that the LSTM model delivers promising performance.

[3]15-Classes: *inactive*; stop: *both-static*, *both-dynamic*, *left-static*, *left-dynamic*, *right-static*, *right-dynamic*; clear: *left-static*, *right-static*; go: *both-static*, *both-dynamic*, *left-static*, *left-dynamic*, *right-static*, *right-dynamic*.

VI. CONCLUSION

We introduced a road traffic control gesture recognition dataset in the context of autonomous driving. Our dataset consists of 3D body skeleton data and gesture category for every time step. To perform gesture classification, we presented eight sequential processing models based on deep neural networks, such as recurrent networks, temporal convolutional networks and graph convolutional networks. Finally, we demonstrated promising performance on real-world sequences, which indicates the representativity for our dataset.

ACKNOWLEDGMENT

Part of the research was conducted within @CITY-AF (Research project No. 19 A 18003 A.), funded by BMWi (Federal Ministry for Economic Affairs and Energy).

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[2] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.

[3] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "The EuroCity Persons Dataset: A Novel Benchmark for Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5 2019.

[4] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *European Conference on Computer Vision*. Springer, 2014, pp. 459–473.

[5] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, *et al.*, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 190–204, 2017.

[6] N. Zengeler, T. Kopinski, and U. Handmann, "Hand gesture recognition in automotive human–machine interaction using depth cameras," *Sensors*, vol. 19, no. 1, p. 59, 2019.

[7] V. Belagiannis, C. Amann, N. Navab, and S. Ilic, "Holistic human pose estimation with regression forests," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8563 LNCS, pp. 20–30, 2014.

[8] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures revisited: Multiple human pose estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 1929–1942, 2015.

[9] V. Belagiannis and A. Zisserman, "Recurrent Human Pose Estimation," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 468–475. [Online]. Available: http://arxiv.org/abs/1605.02914

[10] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[11] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[12] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[13] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[14] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152–165, 2015.

[15] C. A. Pickering, K. J. Burnham, and M. J. Richardson, "A research study of hand gesture recognition technologies and applications for human vehicle interaction," in *2007 3rd Institution of Engineering and Technology Conference on Automotive Electronics*. IET, 2007, pp. 1–15.

[16] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE transactions on intelligent transportation systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

[17] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE transactions on intelligent transportation systems*, 2019.

[18] F. Sachara, T. Kopinski, A. Gepperth, and U. Handmann, "Free-hand gesture recognition with 3d-cnns for in-car infotainment control in real-time," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 959–964.

[19] N. Zengeler, T. Kopinski, and U. Handmann, "Hand gesture recognition in automotive human–machine interaction using depth cameras," *Sensors*, vol. 19, no. 1, p. 59, 2019.

[20] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 1–7.

[21] K. Lindgren, N. Kalavakonda, D. E. Caballero, K. Huang, and B. Hannaford, "Learned hand gesture classification through synthetically generated training samples," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3937–3942.

[22] S. Gupta, M. Vasardani, and S. Winter, "Conventionalized gestures for the interaction of people in traffic with autonomous vehicles," in *IWCTS 16: Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, 2016, pp. 55–60.

[23] C. Ma, Y. Zhang, A. Wang, Y. Wang, and G. Chen, "Traffic command gesture recognition for virtual urban scenes based on a spatiotemporal convolution neural network," *ISPRS International Journal of Geo-Information*, vol. 7, no. 1, p. 37, 2018.

[24] J. He, C. Zhang, X. He, and R. Dong, "Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features," *Neurocomputing*, 2019.

[25] P. K. Pisharady, P. Vadakkepat, and A. P. Loh, "Attention based detection and recognition of hand postures against complex backgrounds," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 403–419, 2013.

[26] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 445–452.

[27] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *2011 IEEE International conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 1114–1119.

[28] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng, "Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3763–3771.

[29] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[30] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 4263–4270, 2017.

[31] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[32] S. Cho, M. Maqbool, F. Liu, and H. Foroosh, "Self-attention network for skeleton-based human action recognition," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 635–644.

[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[34] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[35] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[36] L. Casas, A. Klimmek, N. Navab, and V. Belagiannis, "Adversarial signal denoising with encoder-decoder networks," *arXiv preprint arXiv:1812.08555*, 2018.

[37] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[38] Y. Li, Z. He, X. Ye, Z. He, and K. Han, "Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 78, 2019.

[39] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[40] K. Zou, M. Yin, W. Huang, and Y. Zeng, "Deep Stacked Bidirectional LSTM Neural Network for Skeleton-Based Action Recognition," in *Image and Graphics*, Y. Zhao, N. Barnes, B. Chen, R. Westermann, X. Kong, and C. Lin, Eds. Cham: Springer International Publishing, 2019, pp. 676–688.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. [Online]. Available: http://arxiv.org/abs/1703.06870

[43] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," 11 2018. [Online]. Available: https://arxiv.org/abs/1811.11742