2019 IEEE Intelligent Transportation Systems Conference (ITSC)
Auckland, NZ, October 27-30, 2019

Risky Action Recognition in Lane Change Video Clips using Deep Spatiotemporal Networks with Segmentation Mask Transfer

Ekim Yurtsever*, Yongkang Liu**, Jacob Lambert*, Chiyomi Miyajima***, Eijiro Takeuchi*†, Kazuya Takeda*† and John H. L. Hansen**

***Abstract*— Advanced driver assistance and automated driving systems rely on risk estimation modules to predict and avoid dangerous situations. Current methods use expensive sensor setups and complex processing pipelines, limiting their availability and robustness. To address these issues, we introduce a novel deep learning based driving risk assessment framework for classifying dangerous lane change behavior in short video clips captured by a monocular camera. First, semantic segmentation masks were generated from individual video frames with a pre-trained Mask R-CNN model. Then, frames overlayed with these masks were fed into a time distributed CNN-LSTM network with a final softmax classification layer. This network was trained on a semi-naturalistic lane change dataset with annotated risk labels. A comprehensive comparison of state-of-the-art pre-trained feature extractors was carried out to find the best network layout and training strategy. The best result, with a 0.937 AUC score, was obtained with the proposed framework. Our code and trained models are available open-source[1].**

I. Introduction

Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS) are being developed with the promise of reducing traffic and increasing safety on roads, translating to considerable economic benefits [1]. Automated driving functions categorized as level three and above have already seen some success, typically through lidar and radar perception, but the high cost of these sensing modalities has slowed their integration in consumer vehicles. Moreover, even though remarkable progress has been achieved, vehicles equipped with these technologies are still involved in traffic accidents [2].

In contrast, camera-based solutions to challenging perception tasks are low-cost and increasingly robust. Developments in machine learning, particularly through deep convolutional neural networks (CNNs), significantly increased object detection capabilities [3] and made reliable object tracking achievable [4]. Furthermore, CNNs trained on big datasets became capable of learning generic feature representations. As a result, generalized, multi-task networks were developed [5], as well as end-to-end networks [6], which avoid the need for complex pipelines. Combined with recurrent neural networks (RNNs) and specifically Long-Short Term Memory networks (LSTMs), spatiotemporal relationships can now be modeled for action recognition [7]. In the vehicle domain, velocity estimation with neural networks using a monocular camera was achieved [8]. As such, cameras, under their operational illumination conditions, become a feasible sensing modality for intelligent vehicle technologies. In this work, we propose a novel risk estimation system that uses vision as its sole modality. This increases the implementation possibility of our method, as cameras are inexpensive and readily available in everyday devices such as smartphones.

Fig. 1. Two lane change samples and classification outputs of our method. The average duration of these clips is ∼10 seconds. 0.937 AUC score was achieved with the proposed method. Video samples can be found in our repository[1].

The focus of this study is specifically risk estimation in lane changes. Lane changing is an essential driving action that is executed millions of times on a daily basis. It has been one of the most common pre-crash scenarios, where 7.62% of all traffic accidents between light vehicles can be attributed to it [9]. Only rear-endings occur more frequently, which are primarily due to inattention. On the other hand, understanding a complex driving scene followed by acute decision making is necessary for negotiating a lane change. As such, a tool for unsafe behavior detection during lane change is of paramount importance. Heuristic and rule-based models often ignore the uncertainty of real systems. Forcing handcrafted policies or building deductive theories leads to observing unexpected behavior, which manifests itself as unmodeled dynamics in these approaches. As such, we believe a data-centric, learning based framework is imperative for finding the best explanation of the observed driving phenomena.

Our experiments evaluated several spatiotemporal network architectures on a naturalistic driving dataset consisting of

*E. Yurtsever, J. Lambert, E. Takeuchi and K. Takeda are with Nagoya University, Nagoya, Japan.
**Y. Liu and J. H. L. Hansen are with UT Dallas, Texas, United States.
***C. Miyajima is with Daido University, Nagoya, Japan.
† E. Takeuchi and K. Takeda are also with Tier IV Inc., Nagoya, Japan.
Corresponding author: Ekim Yurtsever, ekimyurtsever@gmail.com
[1]https://github.com/Ekim-Yurtsever/DeepTL-Lane-Change-Classification

860 lane change videos. Individual sequences in the dataset were classified as risky or safe, as shown in Figure 1. We also compared the feature representations of a wide selection of pre-trained state-of-the-art image classification networks.

The major contributions of this work can be summarized as:

- A novel deep learning based driving risk assessment framework with semantic mask transfer is proposed and used for detecting dangerous lane changes.
- Using solely a camera for the task
- Extensive comparison of state-of-the-art deep backbone models with real-world data

The rest of the paper is organized as follows: after reviewing related literature in risk estimation, spatiotemporal classification and transfer learning, we describe the proposed method in Section III. Then, the experimental setting is explained in Section IV, followed by results in Section V.

II. Related Works

A. Risk Studies

Safety is a key factor driving intelligent vehicle technologies forward and an active area of research. Recently proposed ADAS usually attempt to detect and track surrounding objects and decide whether an action is necessary. Despite the successful implementation of such systems, there is no common agreement on the definition of risk. An objective definition based on a statistical probability of collision was proposed in [10]. This *objective risk* framework led to research focusing on vehicle tracking, where motion models predicted the future position of vehicles, allowing risk assessment [11], [12]. On the other hand, risk metrics that consider the indeterministic human element was also proposed. *Subjective risk*, as in the risk perceived by human drivers, was studied as an alternative [13]. The findings of this study indicate that human driving behavior is based on a perceived level of risk instead of calculated, objective collision probabilities. Bottom-up unsupervised learning approaches were shown to be working for extracting individual driving styles [14], but the latent learned representations were not associated with risk due to the nature of unsupervised learning.

Lane change is a typical driving maneuver that can be performed at a varied level of risk, and a significant percentage of all crashes happen due to the erroneous execution of it [9]. Risk in lane changes was studied mostly from the perspective of objective collision risk minimization [11], [15]. A lane change dataset with manually annotated subjective risk labels made it possible to approach this problem from the perspective of supervised learning. The dataset includes ego-vehicle signals such as steering and pedal operation, range information and frames captured by a front-facing camera close to the drivers' point of view. However, previous works on this dataset ignored the monocular camera footage and used ego-vehicle signals [16], [17].

We utilized the video clips of the aforementioned dataset and focused solely on 2D vision in this study.

B. Spatiotemporal Classification

Image-based spatiotemporal classification research primarily focuses on video classification. An active application in this field, closely related to this work, is action recognition, where a sequence of 2D frames must be classified into one of many, some very similar, actions. The widely used UCF101 action recognition dataset [18] features 101 actions such as running or soccer penalty kicks, which are difficult to differentiate when examining individual frames. Another widely used dataset is the Sports1M dataset [19], which features one million videos of 487 classes of sports-related actions.

The spatial relationship of things forms the context of a single image frame. The spatiotemporal context, on the other hand, is constituted by the motion of things, spanned across time in multiple frames. This makes action recognition a more challenging problem than image classification. Furthermore, in the case of a moving data collection platform, such as a vehicle equipped with a camera, distinguishing the local spatiotemporal context (the motion of things *in* the scene), apart from the global context (the motion *of* the scene), increases the difficulty of the problem.

While video classification has a history of using traditional computer vision, the current state-of-the-art is entirely dominated by deep learning approaches. Transfer learning is a staple in these methods: CNNs pre-trained on large image datasets are used as a starting point, then modified for spatiotemporal classification and fine-tuned using action recognition datasets. Early work introduced two network archetypes, one in which spatial and temporal features were extracted simultaneously by a single network [19] and the other which had two distinct spatial and temporal branches followed by fusion [20]. Recently, both 3D CNNs [21] as well as RNN variants, especially LSTMs, have been used to approach this problem [22]. Various 2D CNNs have been shown to produce good input features for temporal networks like LSTMs. Optical flow output from CNNs has also been used as input to LSTMs [7]. Closely related to this work, features extracted from deep CNNs have been used as input to temporal networks [23], [24].

C. Transfer Learning

Transfer learning can be generally thought of as modifying an existing network for some other application. More precisely, given a source domain and learning task, the aim is to transfer the source model's knowledge to a target model, which may have another target domain and task. Transfer learning methods are further classified depending on how the source and target domain and task differ. Inductive transfer learning refers to the case where the source and target tasks are different, whereas in transductive transfer learning, the domain changes while the task remains the same [25].

Transfer learning has been used for diverse applications with varying complexity. It has been used for natural language processing [26], speech recognition across different languages [27], voice conversion [28] and other computer vision applications [29]. The instance segmentation algorithm

Mask R-CNN [5], which is used in this paper, demonstrated the benefits of multi-task transfer learning. It was trained for both bounding box estimation and instance segmentation, yet it was shown to outperform its previous work which focuses on the former [30]. Furthermore, when its knowledge was transferred to the task domain of human pose estimation, it significantly outperformed competing algorithms. This shows the potential of transfer learning for network generalization, which was leveraged in this research.

Inductive transfer learning was used in this work, as the target domain is somewhat similar to the source domain, but the target task is entirely different. Specifically, the assumption tested here is that feature representations learned by deep CNNs trained on large image databases should be transferable to our task and domain: classifying lane change video clips as safe or dangerous. Feature representations obtained from several pre-trained networks were utilized to test this assumption as outlined in Section III-D.

III. Proposed Method

A novel deep spatiotemporal driving risk assessment framework with Semantic Mask Transfer (SMT) is proposed here. The proposed strategy was used for recognizing risky actions in short lane change clips. Furthermore, an exhaustive comparison of state-of-the-art deep feature extractors was carried out to find the best model layout.

A. Problem Formulation

The principal postulation of this study is the partition of the whole lane change set into two jointly exhaustive and mutually exclusive subsets: safe and risky. This proposition is an oversimplification and, depending on the domain, may not suffice. Nevertheless, this dichotomy simplifies problem formulation and enables the employment of state-of-the-art binary video classification methods. All hypotheses that contradict with this postulation are out of this study's scope.

The objective is to classify a sequence of images captured during a lane change into the risky or the safe subset. The temporal dimension of videos can vary depending on the application and the lane change itself. However, a fixed number of frame constitution is assumed in this study. This decision enables the deployment of fixed-dimension network architectures to solve the problem. The classification problem is formulated as follows:

For lane change i, the goal is to find the inferred risk label $\hat{y}_i$, given the sequence of images captured during the lane change $\mathbf{x}_i = (x_1, x_2, \cdots, x_T)$, with the spatiotemporal classification function f.

$$\hat{y_i} = f(\mathbf{x}_i) \tag{1}$$

where T is fixed $\forall$ lane changes and risk label y is encoded as a one-hot vector.

$$y = \begin{cases} (1,0) & \text{for safe lane changes} \\ (0,1) & \text{for risky lane changes} \end{cases}. \tag{2}$$

The spatiotemporal classifier, f, is learned with supervised deep learning models. Extraction of the ground truth, y, is explained in Section IV-A

B. Deep Neural Network Architectures

Besides the proposed framework, spatiotemporal classification with semantic mask transfer (SMT+CNN+LSTM), a significant contribution of this study is the comprehensive experimental analysis and evaluation of the state-of-the-art video classification architectures for the task at hand.

Two different learning strategies were followed in this work. The first one was the conventional supervised deep learning approach: training a deep neural network architecture from scratch with raw image input and target risk labels through backpropagation. In the second approach, transfer learning was utilized for extracting high-level abstract features from the raw image data. After extraction, these features were fed into separate classifiers which were trained using the target risk labels. A wide selection of pre-trained state-of-the-art very deep networks was used as feature extractors in the experiments.

Furthermore, six architecture families were compared throughout the experiments. Details of each are given in the following sections. High-level diagram of the proposed method, labeled as SMT+CNN+LSTM, is shown in Figure 2.

C. Training From Scratch

Deep learning is a popular machine-learning algorithm family. It is widely used especially for computer vision tasks. However, huge amounts of data are required to train deep architectures. Without adequate data, the performance drops significantly. The lane change dichotomy that is introduced here is not a well-established domain in comparison to the standard image classification problem. As such, big datasets that are annotated laboriously such as ImageNet [31] and COCO [32] do not exist for the task. Therefore, a specific lane change dataset was collected and annotated [33]. The scope of this corpus, however, is not on the same scale as the mentioned datasets.

Two different architectures were used for the training from scratch stratagem. CNNs are an integral part of state-of-the-art image classification models. As such, Frame-by-Frame (FbF) classification with CNNs was selected as the baseline here. The baseline was compared against the state-of-the-art spatiotemporal classification architecture; the CNN + Long Short-Term Memory (CNN+LSTM) model.

Frame-by-frame classification with CNNs (FbF CNN)

A CNN architecture with a fully connected softmax final layer was designed as the baseline in this study. The temporal dimension of lane change clips was disregarded in the baseline. In other words, each frame was classified independently as safe or risky.

The architecture is given in shorthand notation as follows: $x_j \rightarrow C(64,5,1) \rightarrow P \rightarrow C(32,5,1) \rightarrow P \rightarrow \text{FC}(1000) \rightarrow \text{Softmax}(2) \rightarrow \hat{y}_j$. Where $C(r,w,s)$ indicates a convolutional layer with r filters, a $w \times w$ window and s stride size. P stands for max pooling layers and FC(h) for a fully connected dense layer with h hidden units. The final layer is a fully connected softmax with 2 classes. In order to train

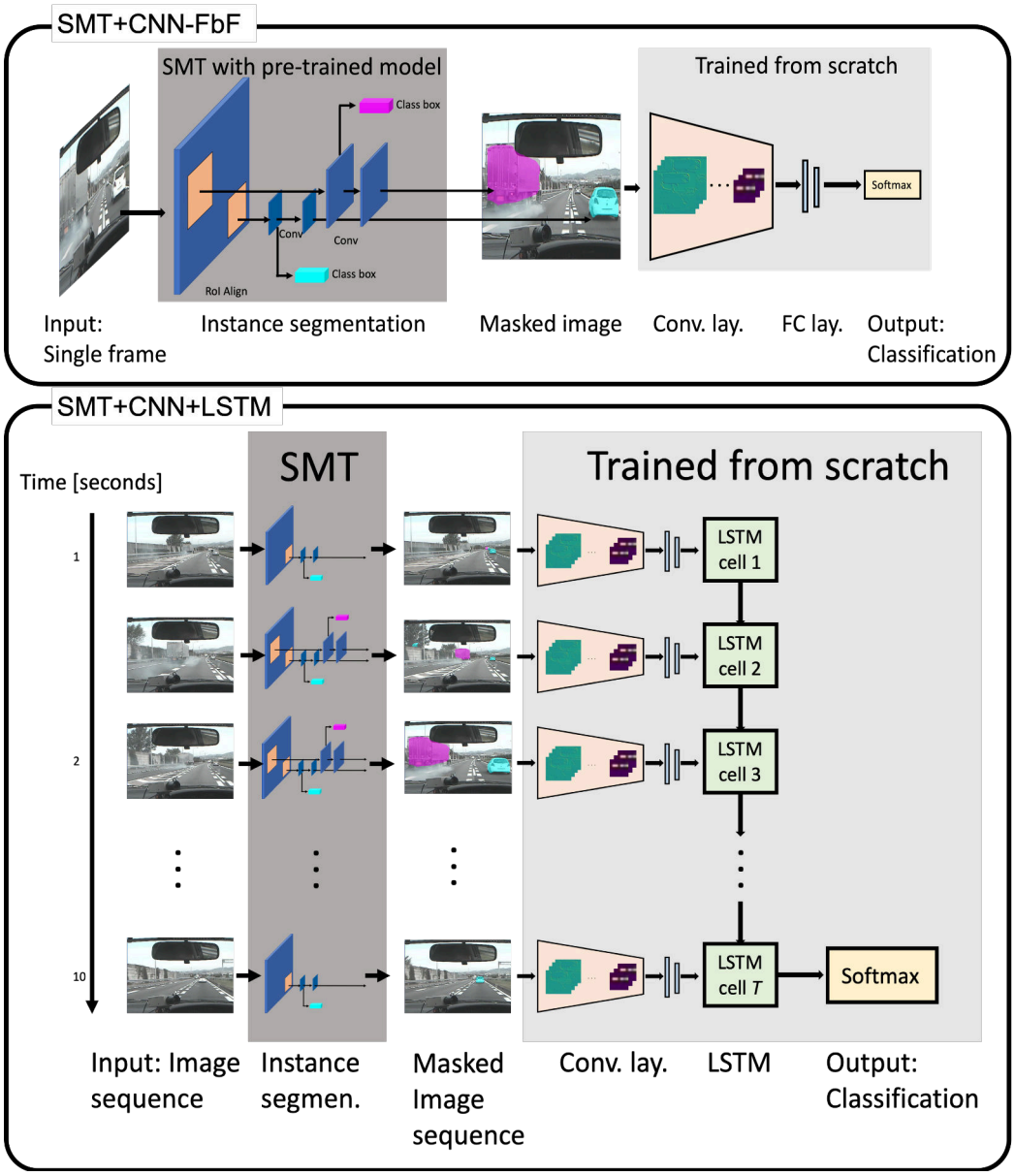

Fig. 2. The proposed framework. SMT stands for semantic mask transfer. Mask R-CNN [5] is used as the mask extractor, and no fine-tuning is done for the SMT part. In this implementation, trucks are colored in magenta and cars in cyan. The average duration of lane change clips is $\sim$ 10 seconds. Each clip is subsampled before processing. Details of frame selection is shown in Figure 3. A temporal composition with contrasting elements can be more useful for relaying semantics of the scene. In our subjective opinion, after a glance, the masked image sequence relays a more striking version of the lane change action than the raw sequence.

this network, risk labels were replicated to each constituent frame correspondingly:

$$\forall x_j \in \mathbf{x},\ y_j = y \tag{3}$$

where x_j is the jth frame of the lane change i.

Spatiotemporal classification (CNN+LSTM)

Video classification is a spatiotemporal problem. Therefore, architectures that consider spatial and temporal aspects of the input data are expected to perform better than the baseline.

State-of-the-art performance for action clip classification was achieved in [7] with a Long-Short Term Memory (LSTM) network where CNN features were fed as inputs of each time step. The distinction between two separate actions such as walking and jumping *might* be easier to detect than a postulated difference such as a risky or safe lane change. However, even though action clip classification is not an identical problem to the lane change dichotomy, it is still relevant and shares the same modality. As such, a similar architecture is proposed here to solve the problem at hand as follows:

$$\hat{y} = f_{\text{LSTM}}(f_{\text{CNN}}(x_1), \cdots, f_{\text{CNN}}(x_j), \cdots, f_{\text{CNN}}(x_T)). \tag{4}$$

The spatial feature sequence $\mathbf{z} = (z_1, z_2, \cdots, z_T)$ of a given lane change was extracted from the raw image sequence $\mathbf{x}$ with the CNN feature extractor f_{CNN}. $\mathbf{z}$ was then fed into LSTM cells to infer $\hat{y}$.

The LSTM network computes the sequence of hidden

vectors $\mathbf{h} = (h_1, h_2, ..., h_T)$ given the input feature sequence $\mathbf{z}$ by iterating the following equations for each timestep t.

$$g_t = \sigma(W_g z_{i,t} + U_g h_{t-1} + b_g), \quad (5)$$

$$i_t = \sigma(W_i z_{i,t} + U_i h_{t-1} + b_i), \quad (6)$$

$$o_t = \sigma(W_o z_{i,t} + U_o h_{t-1} + b_o), \quad (7)$$

$$c_t = g_t \circ c_{t-1} + i_t \circ \tanh(W_c z_{i,t} + U_c h_{t-1} + b_c), \quad (8)$$

$$h_t = o_t \circ \tanh(c_t), \quad (9)$$

where g_t, i_t, o_t are the activation functions of the forget gate, the input gate and the output gate respectively. c_t is the cell state vector, $\circ$ is Hadamard product i.e element-wise product and W, U, b are weight matrices that are learned through training. σ is the sigmoid activation function.

A many-to-one layout was used as only one label is required per lane change. The last output vector of the LSTM, h_T, was fed into a dense layer with a softmax activation function to infer the risk label $\hat{y} = f_{\text{softmax}}(h_T)$. The shorthand notation of the complete architecture is as follows: $\mathbf{x} \rightarrow x_j \rightarrow C(16,3,1) \rightarrow C(16,3,1) \rightarrow P \rightarrow D \rightarrow \text{FC}(200) \rightarrow \text{FC}(50) \rightarrow z_j \rightarrow \mathbf{z} \rightarrow \text{LSTM}(q,20) \rightarrow \text{Softmax}(2) \rightarrow \hat{y}$. D stands for a dropout layer with 0.2 dropout probability. $\text{LSTM}(q,h)$ indicates an LSTM layer with q time steps and h hidden units. q was changed throughout the experiments. Details of the temporal dimension is given in Section IV-B.

D. Transfer Learning

As mentioned earlier, supervised training of very deep networks requires enormous amounts of data. This creates a bottleneck for certain problems such as the lane change dichotomy due to the lack of a big dataset. This issue can be circumvented with the use of models that are pre-trained on big datasets. Even though the target task, classifying a sequence of lane change images as risky or safe, is different from the source task of classifying a single image as one of the thousands of classes of ImageNet [31] dataset, pre-trained state-of-the-art networks can be utilized as feature extractors. In this study, four different transfer learning architectures were compared.

Frame-by-frame classification with feature transfer (FbF FT)

Frame-by-frame classification with feature extraction is accepted as the baseline transfer learning strategy of this study. The method is straightforward: first, the pre-trained very deep network was cut before its final fully connected layer. Then, for each frame x_j, the transferred spatial feature z_j was obtained with the truncated pre-trained network f_t.

$$z_j = f_t(x_j). \quad (10)$$

Finally, the extracted feature z_j was fed into a shallow fully connected softmax classifier that was trained with the lane change data to infer the risk labels. VGG19 [34], MobileNet [35], InceptionResNet [36], NasNET [37], Xception [38] and ResNet [39] were used as feature extractors in the experiments. All of the networks were pre-trained on the ImageNet [31] dataset.

Spatiotemporal classification with feature transfer (FT+LSTM)

The second strategy had the same spatial feature extraction step, but a full temporal network was trained instead of a shallow classifier with the lane change data. The same pre-trained networks used for the FbF FT were utilized again for feature extraction.

In summary, FbF FT and FT+LSTM are similar to the training from scratch strategies, namely FbF CNN and CNN+LSTM. The only difference is the replacement of training of convolutional layers with feature transfer.

Frame-by-frame classification with semantic mask transfer (FbF SMT+CNN)

Multi-task deep networks have become popular recently, especially for vision tasks, in urban driving applications. State-of-the-art multi-task networks YOLOv3 [40] and Mask R-CNN [5] were used for segmentation mask transfer in this study. Both of the networks were pre-trained on the COCO [32] dataset.

The performance of the pre-trained Mask R-CNN can be qualitatively analyzed by inspecting Figure 2. The segmentation masks shown in the figure were obtained for the lane change dataset *without* any fine-tuning or training. It is the out-of-the-box performance of Mask R-CNN trained on the COCO dataset, with our inputs.

A slight post-process modification was done to YOLOv3 in order to obtain segmentation masks. The original network outputs a bounding box and class id for detected objects. For each class, bounding boxes were filled with 0.7 opacity and with a unique color in this study. Mask R-CNN did not need any modification as it outputs a segmentation mask besides bounding box and class id. It is assumed that a high-contrast composition is more useful for discerning distinct elements. As such, bright and unnatural colors such as cyan and magenta were selected as mask colors.

The out-of-the-box performance was good but not perfect. Since the ground truth for segmentation masks are not available for the lane change dataset, fine-tuning was not possible nor any quantitative analysis. However, these masks can still be used for risk detection. The idea is very similar to FbF CNN: first, the lane change frames were converted to the masked images with pre-trained networks. After this step, the same network architecture of FbF CNN was used to infer risk labels.

Spatiotemporal classification with semantic mask transfer (SMT+CNN+LSTM)

The main contribution of this work, SMT+CNN+LSTM, is a novel framework for binary video classification. To the best of authors' knowledge, semantic segmentation masks had not been fed into an LSTM architecture for risk detection in the literature before. The proposed method is shown in Figure 2.

The main hypothesis is that a temporal composition with highly contrasting elements can tell a better story. Qualitative evaluation of this claim can be done by inspecting Figure 2.

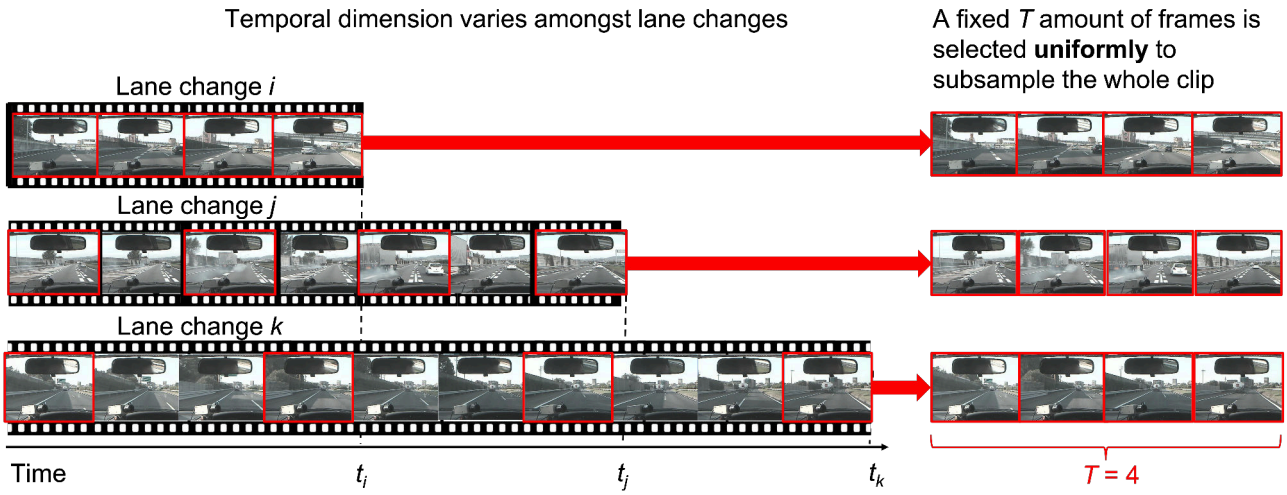

Fig. 3. Subsampling the video clips. A fixed amount of T frames is selected uniformly for each lane change. This design choice enables the employment of fixed frame rate architectures. T is a hyper-parameter of our framework and it affects the classification performance. We tested different values of T throughout the experiments. This point is further elaborated in Section V.

In our subjective opinion, after a glance, the masked image sequence relays a more striking version of the lane change action than the raw sequence. Quantitative analysis is given in Section V.

The starting point of the framework is the masked images whose extraction is described in the previous section. These masked images were passed through convolutional layers to extract abstract features from the *contrasted compositions* created by the mask colors. These high-level features were then fed into LSTM cells to depict temporal relationships. The same CNN+LSTM architecture that is given in Section III-C was used with the only difference being the input type. SMT+CNN+LSTM uses images with overlayed semantic segmentation masks as its sequential input.

IV. Experiments

A. Dataset

A subset of the NUDrive dataset was used in this study. Data was collected with an instrumented test vehicle in Nagoya, Japan. Details of the corpus can be found in [33]. The subset consists of 860 lane change video clips captured by a front-facing camera. Eleven different drivers executed the lane changes on Nagoya expressway. Drivers followed the same route and were asked to keep their natural driving habits while doing lane changes as much as possible. The whole trip of each driver was parsed manually to extract the lane change clips afterward.

The footage was captured with a resolution of 692x480 at 29.4 frames per second. The average duration of a lane change clip is approximately 10 seconds.

Establishing the ground truth: Ten annotators watched the video clips and rated the risk level of each instance subjectively. Annotators gave a risk score between one (safest) and five (most risky) to each lane change. Risk ratings were normalized for each annotator. Then, the normalized scores obtained from ten annotators were averaged to obtain a single score per lane change. The riskiest 5% of the lane change population was accepted as risky while the rest was assumed to be safe. Risky lane changes were taken as the positive class in this binary classification. The final distribution is 43 to 817 for the positive and negative classes respectively.

B. Experimental Conditions and Evaluation Criteria

Temporal dimension length, which is equal to the number of frames fed into the LSTM architecture, was changed throughout the experiments. The number of frames affected the performance significantly. Details of this phenomena are discussed further in Section V.

For each architecture and cross-validation fold, seven different training sessions were run with 5, 10, 15, 20, 50, and 100 frames that were subsampled uniformly per lane change sequence. The average total number of frames per lane change is around 300. The uniform video subsampling is shown in Figure 3.

10-fold-cross-validation was applied all through the experiments. 18 architectures and 6 subsampling options were compared with 10-fold-validation, which totaled in training of 1080 networks.

The lane change dataset is heavily skewed towards the negative class. Accuracy is not a definitive metric under this circumstance. Instead, Area Under the Curve (AUC) was chosen as it is widely used for binary classification problems with large class imbalance. The evaluation focus of AUC is the ability for avoiding false classification [41]. Besides classification performance, inference time is an important criterion. Especially for real-time applications very deep networks can get cumbersome. The main factor that affects inference time is the number of total parameters in an architecture. The evaluation of the experiments with respect to these criteria is given in Section V.

C. Training and Implementation Details

The Adam optimizer was used throughout the experiments with 0.0001 learning and 0.01 decay rate. A batch size of 32 was used on each training run which consisted of 1000

TABLE I

CLASSIFICATION PERFORMANCES ON THE LANE CHANGE DATASET. OUR METHOD, SMT+CNN+LSTM, ACHIEVED THE BEST AUC SCORE.

Architecture	Backbone model	Pre-trained on	# Parameters (millions)	Best T	AUC
FbF CNN	-	-	1.6	1	0.815
FbF FT	VGG19 [34]	ImageNET [31]	144.1	1	0.809
FbF FT	MobileNet [35]	ImageNET	3.7	1	0.779
FbF FT	Inceptionresnet [36]	ImageNET	56	1	0.617
FbF FT	NASNet [37]	ImageNET	89.5	1	0.738
FbF FT	Xception [38]	ImageNET	23.1	1	0.683
FbF FT	ResNet50 [39]	ImageNET	25.8	1	0.861
FbF SMT+CNN	YOLOv3 [40]	COCO [32]	63.6	1	0.854
FbF SMT+CNN	Mask RCNN [5]	COCO	65.8	1	0.853
CNN+LSTM	-	-	0.6	50	0.888
FT+LSTM	VGG19 [34]	ImageNET	143.9	50	0.886
FT+LSTM	MobileNet [35]	ImageNET	3.6	10	0.844
FT+LSTM	Inceptionresnet [36]	ImageNET	56	20	0.5
FT+LSTM	NASNet [37]	ImageNET	89.3	5	0.761
FT+LSTM	Xception [38]	ImageNET	23	50	0.768
Best 3					
FT+LSTM	ResNet50 [39]	ImageNET	25.8	20	0.910
SMT+CNN+LSTM	YOLOv3 [40]	COCO	62.5	50	0.927
SMT+CNN+LSTM	**Mask R-CNN** [5]	COCO	64.8	50	**0.937**

FbF: Frame-by-frame, FT: Feature Transfer, SMT: Semantic Mask Transfer

epochs. Training to validation split-ratio was 0.9 for all cross-validation runs. A categorical cross entropy loss function was employed for all architectures.

The proposed approaches were implemented in Keras, a deep learning library for Python. Our code is open-source and can be accessed from our GitHub repository[2]. The computational experiments took less than a month to finish. A GPU cluster with 6 Nvidia GTX TITAN X was utilized for this research.

V. RESULTS

Table I summarizes the experimental results, where network architectures are shown in the first column. The backbone model column indicates the base transferred very deep network if there was any. Not all architectures used transfer learning, namely FbF CNN and CNN+LSTM. Datasets that the transferred networks were pre-trained on are given in column three. It should be noted again that all architectures were trained with our data for the final classification task. The total number of network parameters for each architecture is shown in column four for assessing the computational load. Lower parameter amount correlates with faster inference time. T, the fixed number of frames, were changed between 5 to 100 for each configuration. The best scoring T in terms of AUC of each row is given in column five. The final column is the AUC score, the main performance metric of this study.

All spatiotemporal architectures with an LSTM layer outperformed their spatial counterparts, except the configuration with the Inceptionresnet [36] base model, which had the lowest performance. These results underline the importance of the temporal dimension. However, a very large dependence on temporal information is also undesired because; it swells the network, increases the input data size and slows the inference time.

[2]https://github.com/Ekim-Yurtsever/DeepTL-Lane-Change-Classification

The best result was obtained with the proposed SMT+CNN+LSTM framework which used a Mask R-CNN [5] semantic mask extractor. We believe this result was due to the masked-contrasted temporal compositions' aptitude for relaying semantic information. The third best result was obtained with an FT+LSTM architecture which used ResNet50 [39] as its backbone model. The rest of the architectures fell behind the top three by a noticeable margin. For example, the proposed SMT+CNN+LSTM's risky lane change detection performance was 25% better than the FbF FT with an Xception backbone.

VI. CONCLUSIONS

Classifying short lane change video clips as risky or safe has been achieved with a 0.937 AUC score using the proposed SMT+CNN+LSTM method.

Our experiments bolster the belief in the adaptive capabilities of deep learning. Transfer learning expands the potential use of trained models. With the increasing availability of open-source libraries and fully trained models with high out-of-the-box performance, new problems can be tackled without tailoring huge datasets for them. The results of this study reinforce this claim.

Promising results were obtained in this work, but only a single driving action, the lane change maneuver, was investigated. In order to parse and assess continuous driving footage, more spatiotemporal techniques should be tested such as feature pooling and 3D convolution in future works. Improving the transfer learning strategies with fine-tuning and utilizing more modalities such as lidar are also amongst our future objectives.

ACKNOWLEDGMENT

This work has been partly supported by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] W. D. Montgomery, R. Mudge, E. L. Groshen, S. Helper, J. P. MacDuffie, and C. Carson, "Americas workforce and the self-driving future: Realizing productivity gains and spurring economic growth," 2018.

[2] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th International Conference on Software Engineering*. ACM, 2018, pp. 303–314.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[4] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 749–765.

[5] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[6] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for Self-Driving cars," Apr. 2016.

[7] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[8] M. Kampelmühler, M. G. Müller, and C. Feichtenhofer, "Camera-based vehicle velocity estimation from monocular video," *arXiv preprint arXiv:1802.07094*, 2018.

[9] W. G. Najm, J. D. Smith, M. Yanagisawa, *et al.*, "Pre-crash scenario typology for crash avoidance research," United States. National Highway Traffic Safety Administration, Tech. Rep., 2007.

[10] G. B. Grayson, G. Maycock, J. A. Groeger, S. M. Hammond, and D. T. Field, *Risk, hazard perception and perceived control*. Crowthorne, UK: TLR Ltd., 2003.

[11] C. Laugier, I. E. Paromtchik, M. Perrollaz, M. Yong, J. Yoder, C. Tay, K. Mekhnacha, and A. Nègre, "Probabilistic analysis of dynamic scenes and collision risks assessment to improve driving safety," *IEEE Intell. Transp. Syst. Mag.*, vol. 3, no. 4, pp. 4–19, 2011.

[12] S. Lefèvre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," *ROBOMECH Journal*, vol. 1, no. 1, p. 1, July 2014.

[13] R. Fuller, "Towards a general theory of driver behaviour," *Accident analysis & prevention*, vol. 37, no. 3, pp. 461–472, 2005.

[14] E. Yurtsever, C. Miyajima, and K. Takeda, "A traffic flow simulation framework for learning driver heterogeneity from naturalistic driving data using autoencoders," *International Journal of Automotive Engineering*, vol. 10, no. 1, pp. 86–93, 2019.

[15] H. Woo, Y. Ji, Y. Tamura, Y. Kuroda, T. Sugano, Y. Yamamoto, A. Yamashita, and H. Asama, "Advanced adaptive cruise control based on collision risk assessment," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 11 2018, pp. 939–944.

[16] E. Yurtsever, S. Yamazaki, C. Miyajima, K. Takeda, M. Mori, K. Hitomi, and M. Egawa, "Integrating driving behavior and traffic context through signal symbolization for data reduction and risky lane change detection," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 242–253, 2018.

[17] S. Yamazaki, C. Miyajima, E. Yurtsever, K. Takeda, M. Mori, K. Hitomi, and M. Egawa, "Integrating driving behavior and traffic context through signal symbolization," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 642–647.

[18] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," Dec. 2012.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[21] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3D ConvNets: New architecture and transfer learning for video classification," Nov. 2017.

[22] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[23] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139.

[24] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 350–359.

[25] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[26] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Association for Computational Linguistics*, 2018. [Online]. Available: http://arxiv.org/abs/1801.06146

[27] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, May 2013, pp. 7304–7308.

[28] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned wavenet vocoder," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6815–6819.

[29] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proceedings of Robotics: Science and Systems XII*, July 2015.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[33] K. Takeda, J. H. Hansen, P. Boyraz, L. Malta, C. Miyajima, and H. Abut, "International large-scale vehicle corpora for research on driver behavior on the road," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1609–1623, 2011.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for Large-Scale image recognition," *Computing Research Repository*, vol. abs/1409.1556, 2015.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[37] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 770–778.

[40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018.

[41] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.