Development and testing of an image transformer for explainable autonomous driving systems

Jiqian Dong, Sikai Chen and Mohammad Miralinaghi
Center for Connected and Automated Transportation (CCAT) and Lyles School of Civil Engineering,
Purdue University, West Lafayette, Indiana, USA
Tiantian Chen
Department of Industrial and System Engineering, The Hong Kong Polytechnic University, Kowloon, China, and
Samuel Labi
Center for Connected and Automated Transportation (CCAT) and Lyles School of Civil Engineering,
Purdue University, West Lafayette, Indiana, USA

Abstract

Purpose – Perception has been identified as the main cause underlying most autonomous vehicle related accidents. As the key technology in perception, deep learning (DL) based computer vision models are generally considered to be black boxes due to poor interpretability. These have exacerbated user distrust and further forestalled their widespread deployment in practical usage. This paper aims to develop explainable DL models for autonomous driving by jointly predicting potential driving actions with corresponding explanations. The explainable DL models can not only boost user trust in autonomy but also serve as a diagnostic approach to identify any model deficiencies or limitations during the system development phase.

Design/methodology/approach – This paper proposes an explainable end-to-end autonomous driving system based on "Transformer," a state-of-the-art self-attention (SA) based model. The model maps visual features from images collected by onboard cameras to guide potential driving actions with corresponding explanations, and aims to achieve soft attention over the image's global features.

Findings – The results demonstrate the efficacy of the proposed model as it exhibits superior performance (in terms of correct prediction of actions and explanations) compared to the benchmark model by a significant margin with much lower computational cost on a public data set (BDD-OIA). From the ablation studies, the proposed SA module also outperforms other attention mechanisms in feature fusion and can generate meaningful representations for downstream prediction.

Originality/value – In the contexts of situational awareness and driver assistance, the proposed model can perform as a driving alarm system for both human-driven vehicles and autonomous vehicles because it is capable of quickly understanding/characterizing the environment and identifying any infeasible driving actions. In addition, the extra explanation head of the proposed model provides an extra channel for sanity checks to guarantee that the model learns the ideal causal relationships. This provision is critical in the development of autonomous systems.

Keywords Explainable deep learning, Computer vision, Transformer, Autonomous driving

Paper type Research paper

1. Introduction

1.1 Background

Motivated by the challenges associated with safety and mobility in the traditional highway environment and spurred by ongoing advancements and opportunities in information and robotics technologies, government agencies and the automobile industry continue to seek guidance on the measurement of performance in the context of the new transportation technologies. As is the case with any new transportation stimulus including technological innovations, it is imperative to assess performance based on a

This work was funded by the US Department of Transportation, Award No. 69A3551747105. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views or policies of the sponsoring organization. An abridged version of the paper was presented at the IEEE ITSC conference in Indianapolis, IN, United States, in September 2021.

Author contributions: The authors confirm contribution to the paper as follows: all authors contributed to all sections. All authors reviewed the results and approved the final version of the manuscript.

Received 14 June 2022
Revised 17 June 2022
Accepted 19 June 2022

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2399-9802.htm**

Journal of Intelligent and Connected Vehicles
5/3 (2022) 235–249
Emerald Publishing Limited [ISSN 2399-9802]
[DOI 10.1108/JICV-06-2022-0021]

carefully designed portfolio of performance measures (FHWA, 2019; Sinha and Labi, 2007; World Bank, 2005). In the context of automated (AV) and connected vehicle operations, performance may be measured from the perspective of the impact type (safety, mobility, privacy, equity, for example), impact direction (costs and benefits) and the affected stakeholder (the transportation agency, road user and the community) (Lioris *et al.*, 2017; Litman, 2014; TRB, 2018, 2019). Unfortunately, the deployment of AV systems in the real world has been severely limited due to various obstacles associated with policy and regulation, infrastructure readiness, technology and so on. For example, a number of key technologies associated with perception and decision processors still have not reached a level of advancement where they can be applied reliably to produce error-free AV systems.

1.1.1 Perception is a key consideration

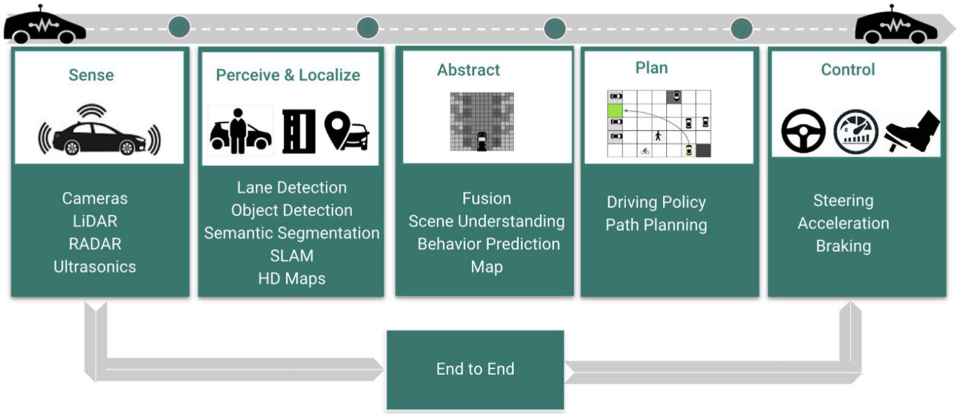
Autonomous driving is a complicated end-to-end system which contains a sequence of subsystems or modules including sensing, perception and localization, abstraction, planning and control (Figure 1), and each module is achieved through the integration of multiple technologies such as sensing, signal processing, data analytics, machine learning, artificial intelligence (AI) and control theory. Of the modules, perception (second block in Figure 1) is generally considered the most vulnerable link in the chain (NTSB, 2019). There are multiple reasons for this. To begin with, the perception module is one of the very initial blocks of the entire autonomous driving process, any error at the perception phase will not only cascade but also be amplified across the subsequent stages. For example, failure in detecting the road participants (i.e. pedestrians, cyclists and neighboring ground vehicles) could be catastrophic because an appropriate evasive maneuver will not be planned in the following phases. This has been the underlying cause of several AV-related fatal accidents in the recent years, including well-known instances of Uber and Tesla vehicle collisions with pedestrians (McCausland, 2019; Yadron and Tynan, 2016).

1.1.2 Application of perception in existing driving systems

Recently, computer vision (CV) based perception technologies have been widely used in multiple applications in driving assistance systems (ADAS) (Horgan *et al.*, 2015; Sowmya Shree and Karthikeyan, 2018), for example, systems for lane detection, traffic sign recognition and forward collision warning. These new features in ADAS have greatly enhanced driving safety and convenience. However, these modules are "scattered" in the sense that they are generally designed to accomplish specific functions in an independent manner. As a result, they do not cooperate with each other and are unable to provide full situational awareness of the driving environment for purposes of autonomous driving. For example, the obstacle detection module can detect only the barriers in the surrounding location but cannot cooperate with the lane marker detection module. Therefore, the obstacle detection module still requires the human brain to fill the gap in such knowledge, and to achieve a comprehensive characterization of the driving environment. Furthermore, these modules in ADAS are currently designed for human driving (where it is required that the human driver is always focused during driving), not AV operations. This means that when developing vision based ADAS, the reliability may be compromised. Therefore, such ADAS systems cannot provide a comprehensive and precise understanding of driving environments, and thus cannot be applied directly to fully AVs. For the perception phases of AV operations, a more sophisticated, integrated cooperative and reliable CV system is needed. In addition, unlike human vision which can quickly identify salient objects and grasp the main semantics in a driving environment, CV models tend to misallocate computation resources towards analyzing areas of the driving environment that may be irrelevant to the driving task (e.g. the background sky and buildings). To alleviate this situation, an appropriate "attention" mechanism to "guide" the CV model to focus only on relevant areas of the driving environment, is needed.

More recently, with the emergence of connectivity devices, perception can be further enhanced by vehicle connectivity (this yields the often-termed "connected autonomous vehicle" [CAV]) as more accurate and direct information can be disseminated through the connectivity devices. It has been postulated that the benefits of combined automation and connectivity will exceed the sum of benefits from these two individually (Ha *et al.*, 2020). In the past few years, many advanced learning based approaches have been applied in the

Figure 1 The end-to-end autonomous driving task



Source: Image from Talpaert *et al.* (2019)

CAV operation in contexts including information fusion and cooperative control (Chen *et al.*, 2021; Dong *et al.*, 2021a; Dong, Chen, Joun Ha *et al.*, 2020; Dong *et al.*, 2020). We duly recognize the coupling of connectivity and automation can accentuate the benefits of the latter. However, there is still a long way to go before achieving full connectivity for all vehicles on road. Therefore, in this paper, we address only the perception tasks of a single vehicle for which we seek to enhance the interpretability of its perception module.

1.2 Literature review

1.2.1 Problems associated with deep learning

In the field of perception and semantic understanding, deep learning (DL) is one of the mainstream technologies which has been used widely in practice. In transportation-related tasks, DL has been extensively adopted in applications including infrastructure management (Hou *et al.*, 2020; Zhuang *et al.*, 2018), traffic prediction (Cui *et al.*, 2019; Liu *et al.*, 2019; Yu *et al.*, 2020a; Zhou *et al.*, 2021), driver behavior modeling (Xing *et al.*, 2021), smart routing systems (Du *et al.*, 2021), smart intersection management (Peng *et al.*, 2021) and traffic incident and duration recognition (Zhu *et al.*, 2021). With respect to autonomous driving tasks, DL models have been applied in every submodule (Figure 1). Specifically, the deep-learning computer vision (DLCV)-based perception models for AV systems are widely researched and have achieved state-of-the-art (SOTA) in various of contexts (Bojarski *et al.*, 2016; Xu *et al.*, 2017; Chen, *et al.*, 2019). Although DL models have been deployed successfully in several real-world applications, the intrinsic drawback, low interpretability, has not been resolved. The low interpretability originates from the black box nature of computations using neural networks. Since the model developer can access only the input and output of the model, the potential weaknesses and drawbacks of DL models are not easily detectable and therefore errors in these models are generally difficult to diagnose. This has exacerbated the problem of user distrust in automation and has further hindered its deployment in safety-critical tasks (Khastgir *et al.*, 2018).

To boost user trust in automation and AI technology, several research efforts have been expended into developing "explainable" AI (XAI) systems. The key motivation and underlying notion of XAI systems is to provide human understandable explanations indicating the rationale used by the AI to make decisions (Doran *et al.*, 2018). This idea has also been adopted into recent transportation-related research efforts. For example, Alwosheel *et al.* (2021) developed an explainable traffic demand prediction model and carried out detailed investigation on how the model provides the predictions (Alwosheel *et al.*, 2021); Bustos *et al.* (2021) applied DL models and provided the interpretability analysis to demonstrate how pedestrian and vehicle safety could be enhanced. In the area of AV system design, researchers have developed explainable (or even, advisable) autonomous driving models (Kim *et al.*, 2020; Kim and Canny, 2017; Xu *et al.*, 2020; Dong *et al.*, 2021b). Here, the "advisable" refers to the model has the ability to process the verbal instructions from human operator and adjust further decisions based on the "advice". These efforts have helped paved the way for enhanced user trust in DL model-driven autonomous driving. Yet still, the model performance (in terms of prediction accuracy and computational cost) can be further improved with an enhanced design of the neural network architecture that imitates human vision. This is the main motivation of this paper. In the subsequent subsections of this introductory section, we discuss two major approaches for developing DLCV-based AV systems, the concept of image-attention based technologies used to imitate human vision. Then we identify the research gaps in existing research and highlight the prospective contributions of this paper.

1.2.3 End-to-end versus pipelined systems

In developing DLCV-based AV systems, there exist two major approaches: end-to-end and pipelined. The former seeks a direct mapping from the raw sensor inputs (including images and 3D cloud points) to the driving actions (including straight movement, left/right turn or slowing down (Bojarski *et al.*, 2016; Chen *et al.*, 2019; Kim *et al.*, 2020; Kim and Canny, 2017; Xu *et al.*, 2017, 2020)). The latter divides the entire system into subsystems (including vision block [Hu *et al.*, 2020; Ku *et al.*, 2019] and decision-generating block [Schwarting *et al.*, 2018; Veres *et al.*, 2011]) and addresses them independently. Theoretically, end-to-end approaches are superior to pipeline approaches because the vision block can get trained to be goal-induced, meaning it is capable of paying more attention to the visual information that is necessary for the ultimate goal. However, the end-to-end approach is generally more complicated and needs relatively deeper networks and larger data sets for training. In addition, because the model is trained from end to end, there are no intermediate results for diagnosis purposes, and this exacerbates the black box behavior. The pipeline approach, on the other hand, is considered more tangible because it can output intermediate results for purposes of human inspection and validation (i.e. object detection bounding boxes). However, the pipeline approach is often suboptimal because training the submodules separately may cause one to lose track of the ultimate goal. Such segregated training can lead to a misallocation of computation resources due to the detection of irrelevant objects or the erroneous neglect of important objects in the driving environment. For example, the detection of objects located beyond the roadway sidewalk will not be beneficial to AVs. However, failure to detect traffic signal colors could be catastrophic. Another limitation of the pipeline approach is that it requires an explicit definition in the manner of cooperation of the two submodules; if the cooperation protocol is ill-defined, the overall performance of the entire model can be jeopardized even if the individual submodules exhibit good performance.

A natural way of integrating the benefits from both end-to-end system and pipelined system is to add intermediate output heads that can generate human understandable results while training the entire system end-to-end for the final goal. For example, by adding "explanation head" to AV systems, the model can simultaneously output both driving actions and corresponding explanations. The explanations provide an opportunity to ascertain whether the correct causal relationships (between the driving environment from the input images and actions) have been learned. They also serve as extra labels (extra loss function) to facilitate the entire training process. Furthermore, from an application perspective, the joint prediction of explanations and decisions could yield additional redundancy to the entire system compared to models that output decisions only. Because causal relationship between explanations and driving decisions can be easily stored as simple rules (e.g. perceiving a yellow light on the traffic signal should result in a slow down or stop decision). If the model fails to predict the consistent decisions with explanations, warnings could be sent immediately to the human operator for the requisite intervention. As the result, this extra setting can help enhance the model interpretability, boost confidence in the model and eventually incentivize AV system manufacturers to adopt the model.

1.2.4 Imitation of human vision using image attention

In recent years, despite the fact that DL models have shown great promise in image processing, human vision remains an incontrovertible benchmark. This is because the human eye possesses a multiresolution structure, namely, peripheral and foveal vision (Xia *et al.*, 2020), and a proper "attention" mechanism to reach a balance in efficiency and recognition accuracy. Peripheral vision is blurry (low resolution) but requires only a short time for processing and has a larger field of vision. Foveal vision, on the other hand, is clear (high resolution) but requires longer processing time and only has a limited vision field. The combination of these two structures guarantees the efficiency because it enables the human to "attend" only to the salient and important regions with foveal vision while the overall information of the scene can be necessarily understood with only peripheral vision. In AV perception, this is also important because the computation cost needs to be minimized as much as possible so that perception and decisions can take place with minimal delay.

In efforts to imitate human vision, visual attention has gradually evolved into a research area of great interest to AV researchers. Recently, a SOTA paper on end-to-end AV systems proposed an object-induced attention mechanism to generate driving decisions with "salient" objects in the scene (Xu *et al.*, 2020). More recently, it has been demonstrated theoretically that a self-attention (SA) based model named Transformer (Zhao *et al.*, 2020) exhibits similar characteristics and performance as the convolutional neural network (CNN) and is also capable of capturing long-range correlations within an image. This is the key inspiration and motivation for the present paper. In subsequent sections of this paper, we demonstrate the efficacy of the Transformer-based model in generating driving actions and explanations for autonomous driving systems.

1.3 Research gaps and main objectives/contributions of this paper

Xu *et al.* (2020) proposed an object-induced attention mechanism that performs an attention over the detected objects and uses only the relative objects to generate driving actions and explanations. More specifically, their model uses a "selector" structure to "crop" the fused regional features. This fused regional feature is generated by stacking the regional features that are computed using Ren *et al.*'s (2017) FasterRCNN's region proposal network (referred to as the local branch) and the raw overall feature map for the entire image (referred to as the global branch). To conduct feature selection, the selector assigns a score to each region proposal to measure its relative importance and identifies the k regions with k-highest scores to compute the driving actions and explanations. That is, during the training process, the model implicitly strives to learn a metric to weigh the regional features based on the semantics and their relative contribution to the driving decisions.

Even though their model demonstrated satisfactory performance in predicting the actions together with explanations, there is still good reason to consider this attention mechanism as suboptimal with certain shortcomings. The shortcomings arise due to three reasons. First, the ablation study results in Xu *et al.*'s research depict a baseline model with only "global" branch can exhibit performance similar to the full model (which integrates the "global" and "local" features of the image). This indicates that the global features (overall information of the image) are more important and can overwhelm the contribution of regional features (the recognized objects in the scenario). This phenomenon is consistent with the intuition that when generating high-level driving actions (move forward, turn left/right, or stop), human drivers tend to use peripheral vision because only an approximate characterization of the driving scene is needed. For example, if there is an obstacle in the driving scene, the driver only needs to roughly see it (with peripheral vision) and can quickly eliminate the erroneous action of driving towards the obstacle's direction before clearly perceiving its details such as shape and color. However, the object-induced attention as described in the research is more consistent to foveal vision because the prediction head of the model can "only" rely on those highly detailed cropped-out regions. Secondly, this attention mechanism largely depends on the object detection module (region proposal network) which has been pretrained to assign greater focus on "object-containing" regions and ignoring "non-object" (background) regions. However, for driving tasks, these "non-object" regions may contain vital information such as lane markers and drivable areas. Therefore, building the model on top of the method based on object detection could be suboptimal. Third, as the number of selected regions k is part of the model parameter, it requires large number of experiments to determine its value. If k too small, this "hard" selection mechanism will inevitably create a bottleneck to restrict the information flow in the model. For example, selecting only k regions will restrain the model flexibility, especially in the cases when there exist more than k pivotal regions (the regions that require the model to attend to achieve full understanding of the scenario). If k too large, the computation resources will be wasted, and the extra information could rather impair the model performance because the noise level is high.

To overcome these three shortcomings, we propose a global soft attention (GSA) mechanism which imitates the peripheral vision capability of the human eye and uses the global features of the image. Overall, the model "softly" fuses the information from each region inside the image using Transformer model. The Transformer model is adopted here because a number of research studies have demonstrated that, compared to CNN, Transformer releases the constraints of generating visual features only based on local regions (Zhao *et al.*, 2020). This makes the model capable of possessing a "broader" horizon and capturing regions that not only are much wider compared to the traditional CNN kernel but also facilitate analysis of correlations within the image. This issue is further discussed in the results section of this paper.

In summary, the main contributions of this paper are threefold:

1 developed an end-to-end explainable DLCV model to generate driving actions with explanations;
2 proposed a new DL architecture with a novel visual attention mechanism using the Transformer model to achieve SOTA with significantly superior performance and lower computational cost compared to the benchmark model; and
3 conducted multiple experiments in a variety of settings to evaluate the importance of information (global vs regional) and the attention mechanism (hard vs soft) in the high-level driving decision-making process.

From the perspective of practical application, the proposed model can enhance human trust in DLCV-based autonomous driving system for both AV users and AV system developers. For AV users, on the one hand, the driving decisions and explanations can be presented to the user simultaneously showing the corresponding causal relationship at the initial deployment stage of autonomous vehicles. On the other hand, such system can perform as a “whistle” to send out instant warnings to the driver or require human intervention if there exist inconsistency between any two predictions. This could lend additional safety redundancy to the entire AV system and thereby boost human trust and acceptance of automated driving systems. For developers, such explainable system is helpful in system debugging because it can output human-understandable outputs, identify potential flaws of the existing system and identify directions for future improvements. Therefore, the concept of “explainable” models is beneficial to the entire AV industry from perspectives of both the user and the manufacturer.

2. Methods

The proposed model, as well as all the baseline models introduced in Section 3 share the same structure containing three blocks, namely, Feature Extractor, Attention Module and Decision/Reason Generation (Figure 2). As its name suggests, the feature extractor is used to generate the low-level feature embeddings from the raw image, which contains image preprocessing (i.e. normalization, reshape) and a pretrained backbone CNN model. On top of Feature Extractor, Attention block solves the problem of information fusion and feature selection. For the proposed model, the attention is achieved by correlating the features from each spatial location using “self-attention” mechanism to compute the attention weights and using these attention weights to either “amplify” or “filter out” the features. The final block takes in the attended feature map and conducts two separate multiclass classification tasks for generating both driving decisions and corresponding explanations. The entire model integrates the three blocks and is trained end-to-end with the aggregated loss function for both predictions. The rest of this section of the paper explains each block in detail.

2.1 Feature extractor

The proposed model used the Feature Extractor block as the global feature extractor to acquire overall features of the image instead of specifically focusing on object-containing regions. To this end, as shown in Figure 3, the module first preprocesses the image (resize and normalization) and then computes the visual features using pre-trained CNN models. In this work, we experimented with two classic backbone CNN models, namely, Resnet50 (He *et al.*, 2016) or Mobilenet_v2 (Sandler *et al.*, 2018). The evaluation and selection of these two models was based on the tradeoff between computation cost and accuracy. Mobilnet_v2 is designed to run on mobile devices, which has much higher computational speed due to its use of a smaller number of parameters, while the Resnet model are much deeper in structure and is believed to represent the state of the art in image feature generation. In addition, as the target for this project is to examine the performance of the upper stream architecture (attention module) and the feature importance, the pretrained Feature Extractor module is frozen in both training and testing time.

2.2 Transformer

The feature map obtained from the Feature Extractor contains the information of the entire image, which is then fed to the Transformer to perform “global soft attention” (Figure 4).

More specifically, the output from the previous block is a 3D tensor of shape ($h \times w \times f$) where f is the feature dimension of each spatial location. The variables h and w represent the height and width, respectively, of the feature map. Then, the two spatial dimensions (h and w) are flattened, and the output 2D feature map $X \in \mathbb{R}^{s \times f}$ is treated as a sequence of input with a sequence length of $s = h \times w$. The basic building block of the Transformer model is referred to as the multihead self-attention (MHSA) layer. MHSA represents a parallel computation of SA which measures the “similarity” between two inputs using their dot product. Initially, SA establishes three representations: key, query and value (K, Q and V) with three distinct linear layers: $K = XW_K$, $Q = XW_Q$ and

Figure 2 Architecture of the proposed model

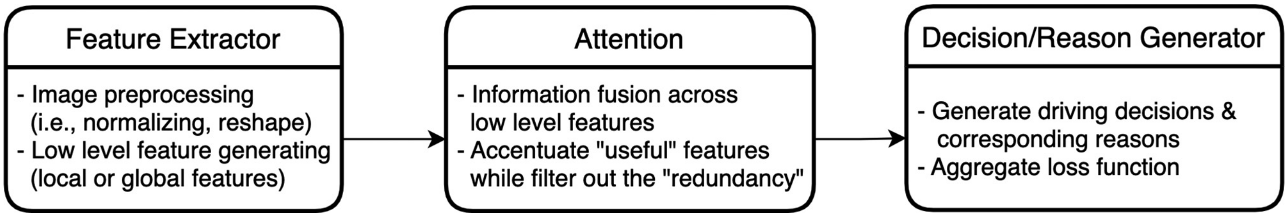

Feature Extractor
- Image preprocessing (i.e., normalizing, reshape)
- Low level feature generating (local or global features)
Attention
- Information fusion across low level features
- Accentuate "useful" features while filter out the "redundancy"
Decision/Reason Generator
- Generate driving decisions & corresponding reasons
- Aggregate loss function

Figure 3 Global feature extractor

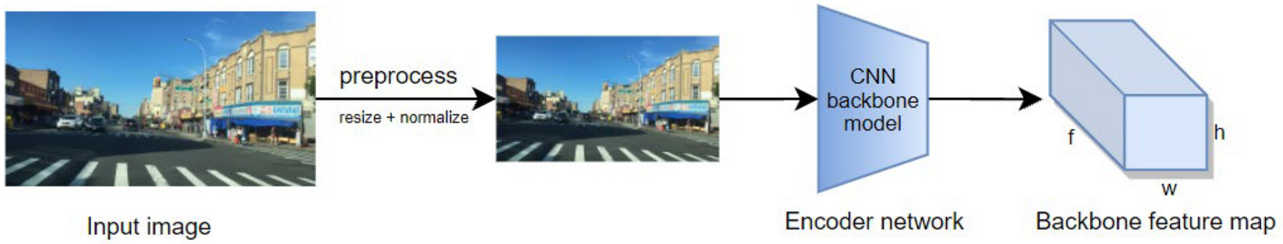
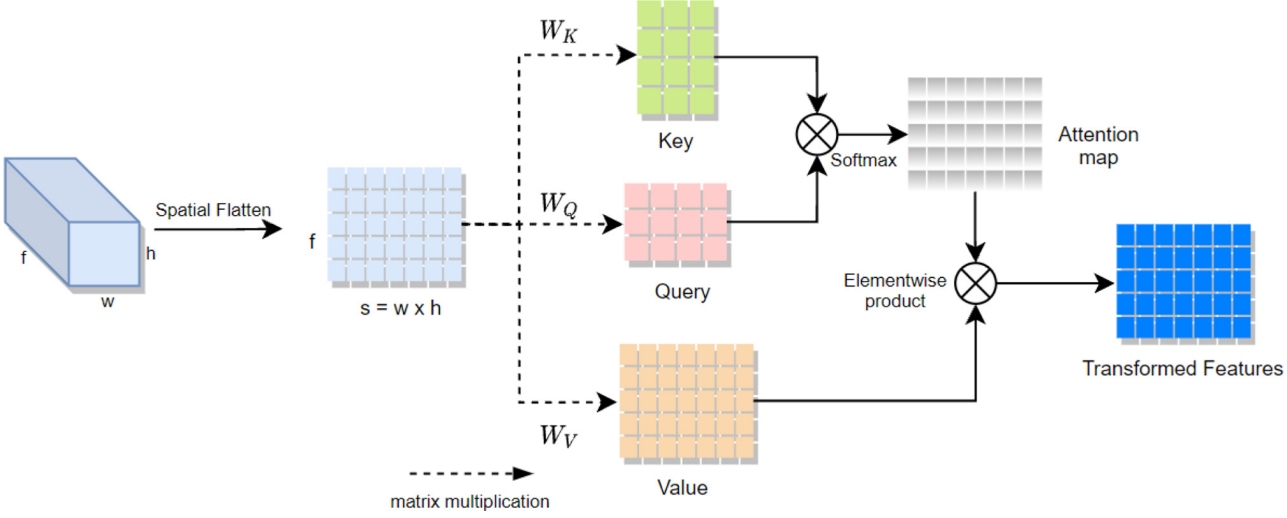


Figure 4 Self-attention (SA) layer (single head)



$V = X^T W_V$, with W_K, W_Q and W_V as their respective weights, $K, \ Q \in \ \mathbb{R}^{s \times d_k}$ and , $V \in \ \mathbb{R}^{s \times d_v}$). It is then possible to generate a matrix of the attention score (a) (which measures degree of correlation between regions) by determining the dot product of the query and key, and then softmax normalization as shown in equation (1):

$$a = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \in \ \mathbb{R}^{s \times s} \tag{1}$$

where d_k is the dimension of K and Q. This attention mechanism enables the output embedding for each spatial location contains not only the information of the spatial location itself but also important information from other spatial locations. The attention scores serve as the fusion weights for generating the attended feature maps. The output of the SA is the multiplication of the value (V) and the attention score, and this completes the computation for single head [equation (2)]. Then the MHSA layer is simply the parallel version of SA which simultaneously computes multiple SA by concatenating all the heads [equation (3)]:

$$head = Attention(K, Q, V) = aV \tag{2}$$

$$MHSA(X) = concat(head_1, \ \ldots, head_h) \ \ W_{out} \in \ \mathbb{R}^{s \times d_{out}} \tag{3}$$

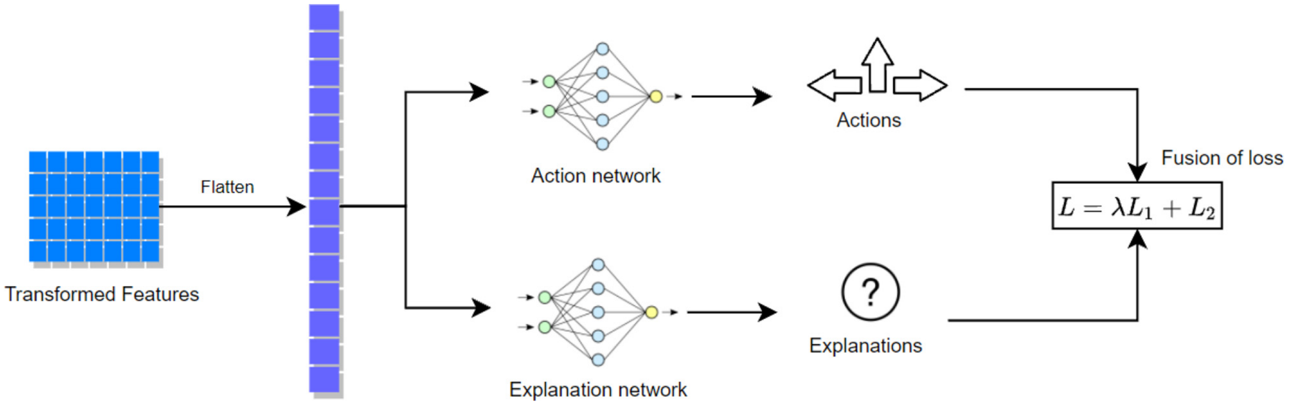
where $W_{out} \in \ \mathbb{R}^{d_v \times d_{out}}$ represent the weights for the final output linear layer which is applied for fusing the results from multiple heads and h is the number of heads computed in parallel. Compared to single head, using MHSA enables each head to simultaneously focus on different tasks and can attend to regions with different ranges. This manipulation can greatly enhance the model's flexibility and generalization power. The final output from the Transformer (MHSA layer) maintains the same spatial dimension as the input feature map X. However, each spatial location contains the "fused" information from this location itself and other regions based on the automatically computed correlation. For demonstration purposes, we use a single MHSA layer in this paper.

2.3 Decision/reason generator

As shown in Figure 5, the Decision/Reason Generator block is a standard multitask classifier containing two branches for generating driving decisions and explanations, respectively. It takes the output feature map from MHSA block as input and performs two separate classifications.

This is achieved using two independent neural networks: the action network and the explanation network. The former has four output classes that represent each driving decisions (going straight, stop/slow down, turn left, turn right) while the latter has 21 output classes for the corresponding explanations. As for the detailed architecture, we use the same structure with fully

Figure 5 Process of the Decision/Reason Generator



connected (FC) layers for both the networks [*Dense*(128) + *Dense*(128) + *Dense*(# *of outputs*)]. Finally, the model is trained end-to-end, following the classic multitask learning manner that aggregates the two losses (driving action loss L_A and explanation loss L_E). This setting requires the model to simultaneously learn to generate the decision and explanation, thus the corresponding causal relationship between these two losses [equations (4) and (5)] can be learned implicitly:

$$L_A = \sum_i^4 L(\hat{A}_i, A_i); L_E = \sum_i^{21} L(\hat{E}_i, E_i) \quad (4)$$

$$L = \lambda L_A + L_E \quad (5)$$

where λ is the weight parameter for tuning the tradeoff between the two losses. Based on the experiment in Xu *et al.* (2020), when $\lambda = 1$, the model yields the best performance in terms of both action and explanation prediction. In this paper, we adopt this result from that study, and therefore, use a $\lambda = 1$ value of 1, because the explanation and driving decision should be equally weighted. The "4" and "21" in equation (4) refer to the total number of actions and explanations, respectively, in the dataset, which will be explained in detail in Section 3 of this paper. $L(s^2,s^2)$ represents the binary cross entropy loss defined in equation (6), where y *and* $\hat{y}$ represent the true label and the model prediction, respectively:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p(\hat{y}_l)) + (1 - y_i)\log(1 - p(\hat{y}_l)) \quad (6)$$

3. Experimental settings

Figure 6 presents the model components for the baseline models.

3.1 Data set

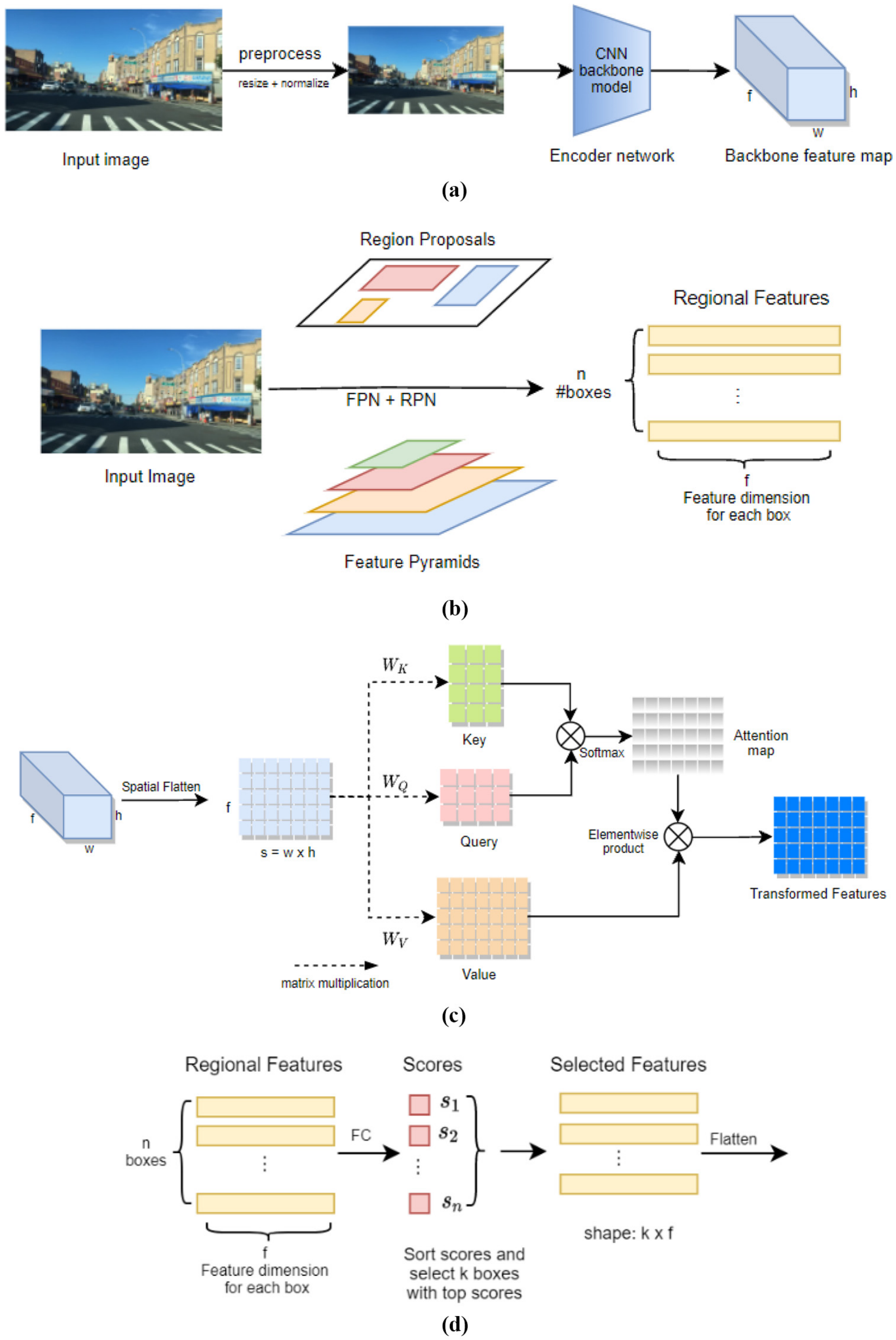
The study trained and evaluated the models described above, using Xu *et al.*'s (2020) BDD Object Induced Actions (BDD-OIA) data set. The BDD-OIA data set extended the original BDD-100K data set (Yu *et al.*, 2020b) by labeling each frame individually with driving actions and explanations. The actions refer to high-level feasible driving maneuvers that can be undertaken by the driver at any specific time step: move forward, stop/slow down, turn left and turn right. The explanations are associated with the actions and are summarized into 21 classes. Figure 7 illustrates the example image and labels. Table 1 presents the labels for actions and explanations. The model is developed with a training set of 16,082 images, a validation set of 2,270 images and a test set of 4,572 images.

3.2 Baseline models and setups

As mentioned in the Introduction, two key technical motivations for this study are to evaluate the relative importance between global and regional information and test different attention mechanisms (hard vs. soft attention). Therefore, we compare our GSA model with several baseline models including: regional hard-attention (RHA), regional soft-attention (RSA) and global no-attention (GNA) model:

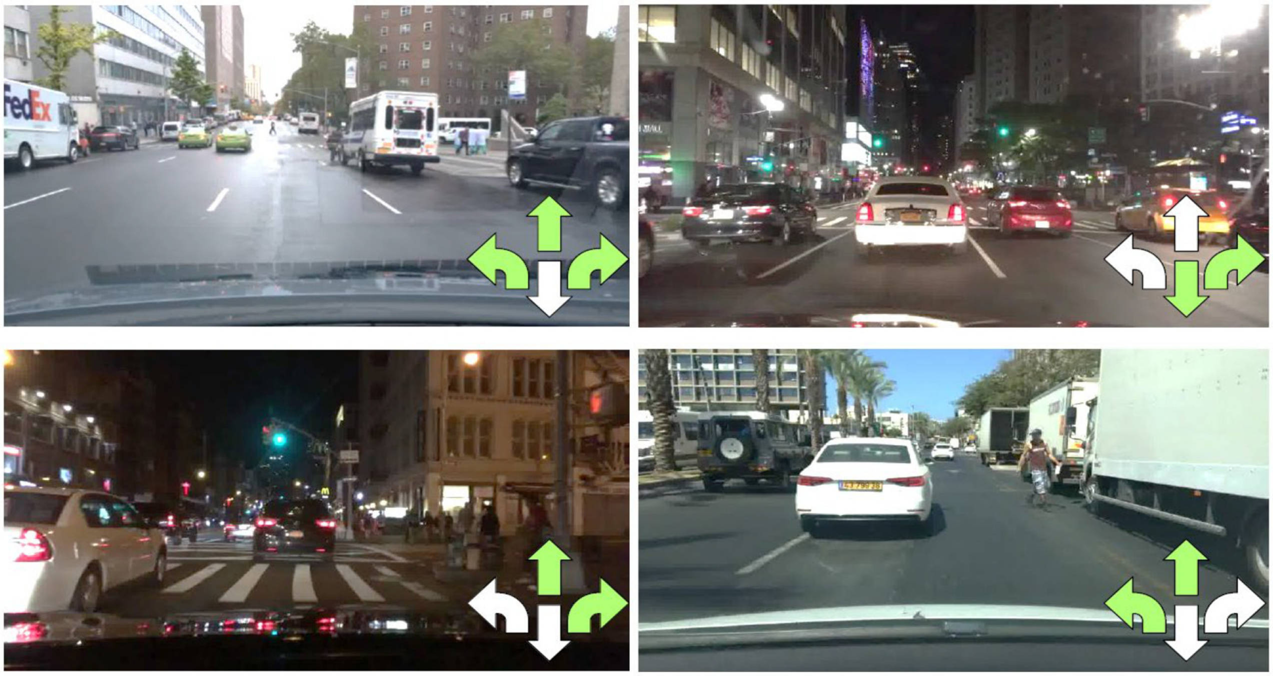
- The RHA model is similar to the object-induced attention model proposed in a benchmark study (Xu *et al.*, 2020) albeit with the local branch only (we did not reimplement the model in the benchmark study but compared its results with ours in the result section). This model uses Faster RCNN with Feature Pyramid Network (FPN) as feature extractor as shown in Figure 6(b), followed by a RHA module which uses a FC layer to compute a score for each region proposal and select only top-k objects (based on the scores) for generating actions [as shown in Figure 6(d)]. In this research, we trained two models with $k = 5$ and $k = 10$. We keep the local branch only to test the importance of regional information for the overall driving decision and explanation generation.
- The RSA model uses the same soft attention mechanism (Transformer) as the proposed GSA model [MHSA block in Figure 6(c)], but the attention is conducted over the region proposals instead of the global features. It uses the same FasterRCNN (FPN) as Xu *et al.* (2020) to generate the regional features [acquired from Figure 6(b)]. This model mainly serves to compare the performance between soft attention and hard attention. In addition, we trained two RSA models with five and eight heads.
- The GNA model mainly serves as an ablation study to our GSA model. It uses the same global features [generated from Resnet/Mobilenet backbone with Figure 6(a) structure] as GSA. However, a vanilla FCN having parameters similar to those of the MHSA block replaces the Transformer structure (MHSA) block.

Figure 6 Structures for the baseline models



Notes: (a) Global feature extractor; (b) regional feature extractor; (c) soft attention; (d) hard attention (only for regional features)

Figure 7 Examples of ground-truth images and decisions from the BDD-OIA data set



Source: Xu *et al*. (2020)

Table 1 Actions with explanations in BDD-OIA

Actions	Explanations
Move forward	Traffic light is green Follow traffic Road is clear
Stop/Slow down	Traffic light is red Traffic sign Obstacle: car Obstacle: person Obstacle: rider Obstacle: others
Turn left	No lane on the left Obstacles on the left lane Solid line on the left On the left-turn lane Traffic light allows Front car turning left
Turn right	No lane on the right Obstacles on the right lane Solid line on the right On the right-turn lane Traffic light allows Front car turning right

4. Results

The training is conducted on one NVIDIA Quadro RTX-6000 GPU, which has 24 GB RAM. All the models are trained using batch size b = 10, the stochastic gradient descent method with an initial learning rate α = 0.001, and a learning rate decay of 10^{-4}. All the models are trained for 40 epochs (64,080 batches). Figure 8 presents the corresponding training curves (training loss vs number of steps) for our proposed GSA model and all the baselines mentioned above. From the training curve, it can be observed clearly that the two proposed GSA models converge much faster and can achieve much lower training losses compared to the baselines.

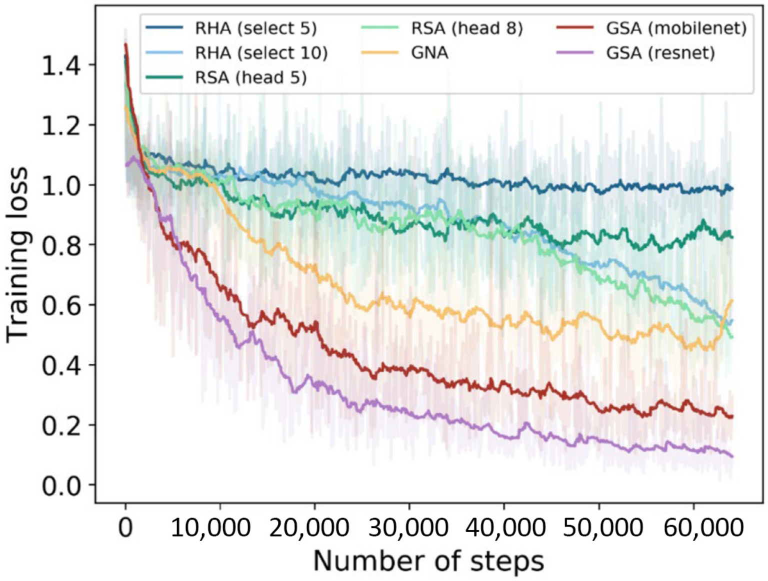
4.1 Quantitative evaluation

We performed the same prediction task (actions and explanations prediction) as the benchmark model presented in previous research (Xu *et al.*, 2020), and we evaluated the proposed model using evaluation metrics similar to those in recent literature. For both decision and explanations, two versions of the F-1 score were used: the overall F1 score, $F1_{all}$ (the F1 score calculated over all the predictions), and the mean in-class F1 score, $mF1$, a metric typically used where the data are unbalanced. A model with a high F1 score indicates that the model has higher recall and higher precision. Equation (7) presents the calculation of the $F1_{all}$ score:

$$F1_{all} = \frac{1}{|A|}\sum_{j=1}^{|A|} F1\left(\hat{A}_j, A_j\right) \tag{7}$$

where A_j = true label (representing an explanation or action), $|A|$ = total number of predictions, $\hat{A}_j$ = predicted value. In the data set, there exist a greater number of instances associated with the "going-straight" action compared to the "turn-left" action; in other words, the data set is unbalanced. For this reason, equation (8) was used to calculate the F1 score for each

Figure 8 Training curve for all seven models (proposed GSA and baselines)



predicted class, and the $mF1$ value was calculated as the mean of all the F-1 scores:

$$mF1 = \frac{1}{C}\sum_{j=1}^{C}\sum_{i=1}^{n} F1\left(\hat{A}_i^j, A_i^j\right) \tag{8}$$

C is the number of predicted classes (4 for actions, 21 for explanations), n is the total number of points in the test data set. The detailed performance in terms of actions and explanations prediction is listed in Table 2.

Apart from the prediction performance, another important aspect of evaluating the model is the computation complexity. We document the number of trainable parameters and the total training time for 40 epochs on the training data set in the final two columns of Table 2. Compared to the regional model (RHA and RSA) which takes more than 10 hours of training, the global models (GNA, GSA) require only one-third of computation resources. The combined results from both loss curve (Figure 8), the performance and computational cost in Table 2 indicate that even with fewer parameters and much shorter training time, the proposed GSA model achieves lower training loss and yields higher performance. This indicates that the global attention is generally superior to regional attention in predicting driving-related actions and explanations.

4.3 Qualitative evaluation

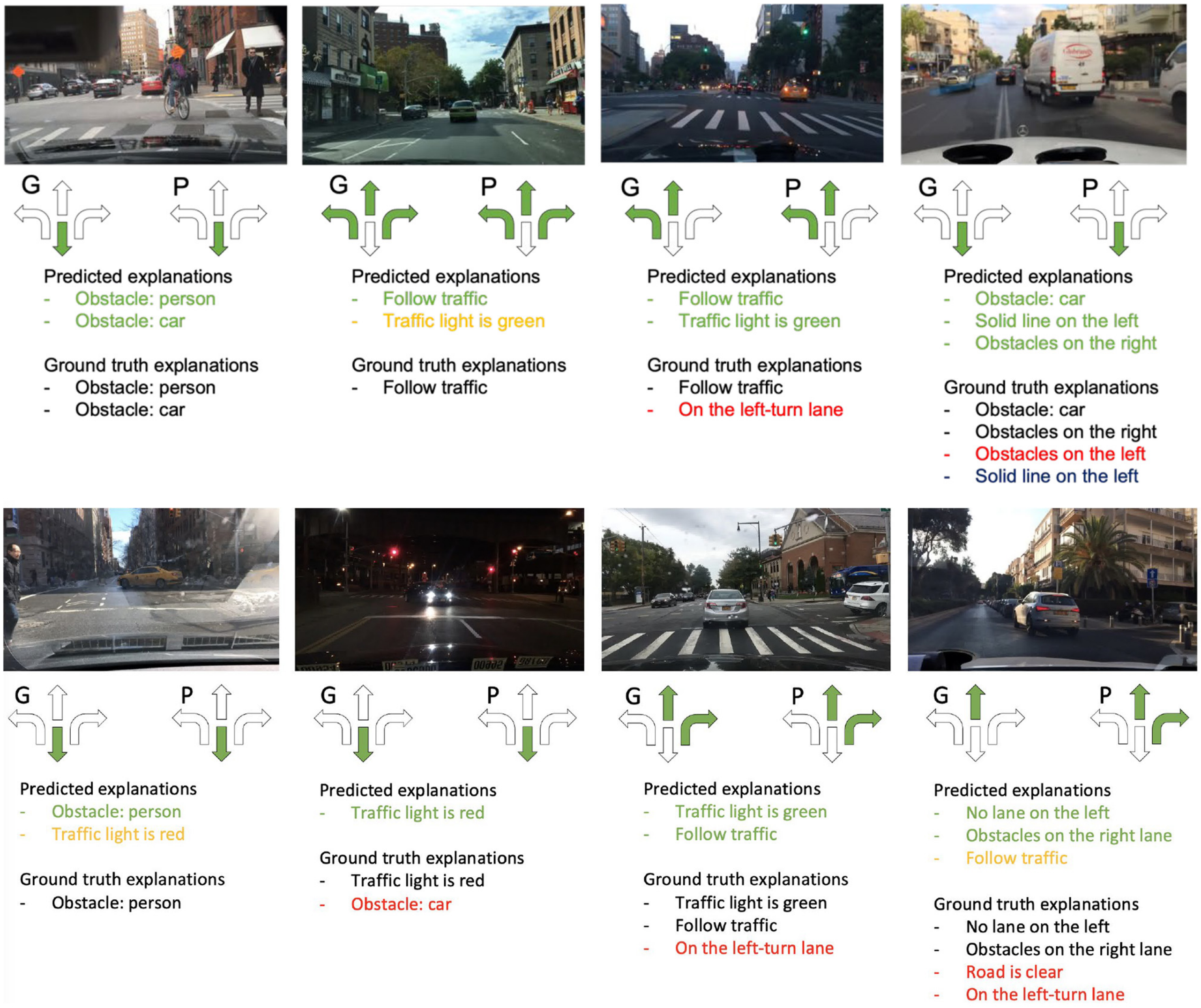
Figure 9 presents the examples of the predictions (generated from GSA-Mobilenet) on the test set. Regarding the action prediction, "G" stands for ground truth and "P" for model prediction. Regarding explanations, green color indicates true

Table 2 Model performance and complexity of proposed model* and the baselines

Attention mechanism	Model	Decision mF1	Decision $F1_{all}$	Explanation mF1	Explanation $F1_{all}$	# of trainable parameters	Training time (40 epochs)
Regional Attention	Xu *et al.* (2020)	0.718	*0.734*	0.208	0.422	–	–
	RHA (5 obj)	0.572	0.494	0.482	0.047	11.04M	10 h, 28 min
	RHA (10 obj)	0.565	0.495	0.499	0.123	21.53M	10 h, 30 min
	RSA (5 heads)	0.595	0.476	0.506	0.127	21.22M	10 h, 32 min
	RSA (8 heads)	0.608	0.542	0.554	0.330	20.75M	10 h, 42 min
Global no attention	GNA (resnet)	0.706	0.660	0.561	0.352	26.10M	3 h, 10 min
****Global soft attention***	*GSA (resnet)*	*0.750*	*0.729*	*0.644*	*0.525*	*24.08M*	*3 h, 15 min*
	GSA (mobilenet)	*0.746*	*0.718*	*0.642*	*0.531*	*2.61M*	*2 h, 53 min*

Notes: From Table 2, the following conclusions can be made: The proposed global soft attention (GSA) models outperform all the baselines with a significant margin, particularly with regard to explanation prediction. Global features are generally more useful compared to regional features, even where the vanilla model (GNA) is used without any special attention mechanism. Soft attention is superior to hard attention even in cases where only regional information is available. With regard to the feature extractor, using Mobilenet_v2 has comparable predictive performance compared to Resnet50 but saves a significant amount of training time. Increasing the number of heads can generally enhance the performance of soft attention models. The superiority of the model is: GSA > GNA > RSA > RNA. This also matches the training loss curve (Figure 8) in terms of the final loss and convergence rate

Figure 9 Example predictions



Notes: Regarding action prediction, "G" stands for ground truth and "P" for model prediction. Regarding explanations, green color indicates true positive, yellow indicates false positive and red indicates false negative

positive, yellow indicates false positive and red indicates false negative. The predictions are done for eight images chosen randomly from the test data set. By inspecting the scenarios and the predictions, the model predictions can correctly predict the driving decisions with high accuracy while generating the corresponding explanations. The generated explanations match the decisions and driving scenarios in most cases. It is worth noting that from the first image, the model predicts an extra explanation indicating the traffic light is red. Even though this is not included in the label, by inspecting the image, we can see the model is in fact making the correct prediction as it does exist the red traffic light in the right of the image. We also notice that there exists some inconsistency in the labels of the dataset, for example, the third image labels the vehicle with turn right decision even if it has the explanation indicating the vehicle is on the left turning lane. This explanation is not predicted by the model since it does not match the correct causal relationship. Furthermore, by inspecting more examples, we notice this inconsistency does not greatly impair the performance of the model and does not impact the comparative analysis between the proposed model with all other baselines.

4.4 Discussion

4.4.1 Attention mechanism: soft > hard

The soft attention (Transformer in this work) is superior to the hard attention (score-based selection) because the former is capable of fusing individual pieces of information in the image based on their individual contributions to the ultimate driving goal (maneuver) rather than simply picking the more important regions. The latter inevitably creates a "bottleneck" to the information flow path and therefore leads to nonconsideration of some information that could be useful to the driving actions. Furthermore, as the regional hard attention "crops" the regions, the correlation between the objects as well as the "relativity" among the image are eliminated. For example, after "selection" operation [Figure 6(d)], obstacles located farther away could have the same representation of the obstacles located close by, then the model cannot know which one is closer. This will increase the ambiguity to the downstream decision/explanation generation block. On the other hand, the soft attention can learn the correlation and compute a "soft" fusion of all the features using the attention map.

4.4.2 Feature importance: global > local

The global features are superior to regional features due to the inherent nature of driving decisions. For generating high-level actions (e.g. move forward, stop/slow), the acquisition of an overall characterization of the roadway scene is more essential compared to the recognition of every single object and computation of their bounding boxes. Therefore, even the GNA baseline can yield superior performance compared to regional attention models built on top of object detection models. In addition, despite the GSA models are not equipped specifically with object detection block in the architecture, the explanations predicted still contains the information of local regions. For example (Figure 9, column 1), the model can still identify the red traffic lights, persons and vehicles obstacles even if these objects occupy only a small proportion of the image. Therefore, based on our experiment results, it is still safe to conclude the global attention (Transformer) mechanism will not neglect the local regions.

4.4.3 Transformer is useful in feature fusion

The Transformer-based models (the two GSAs) outperform the GNA because their MHSA structure can capture long-range correlations within an image. Compared to classic CNN based methods which can capture only the local region correlations due to the fixed size of convolution kernels in each layer, the Transformer-based models enable information fusion over the entire image. This long-range correlation is typically crucial for driving decisions because there exists a "relativity" correlation within the image. For example, "left" is relative to the "right"; therefore, to generate the decision of "turn left," the model needs to understand which part of the image depicts the "left region." Because the cameras are not always facing the same direction as the movement direction of the vehicle, the ratio of "left region" to the entire image keeps changing. Therefore, the model has to understand "left" and "right" relatively from the scene context, which can only be achieved with Transformer based model by simultaneously attending to multiple regions. This entire mechanism is analogous to the peripheral vision of the human eye as human drivers generating driving actions (quickly looking at multiple regions and then making driving decisions instantaneously without clearly seeing each individual object in the region) (Wolfe *et al.*, 2017; Rosenholtz, 2016).

4.4.4 Causal relationship is correctly learned

One of the most salient problem for the existing end-to-end DLCV-based autonomous driving system is that whether the model has truly "understood" the driving scenario remains uncovered to human even if the prediction of driving decisions is correct. In our settings, we "force" the model to explicitly to understand the driving environment by injecting a second loss function (through joint prediction of explanation) as these explanations are the human understandable descriptions to the driving scenario. From Figure 9, it is clearly shown that the model can correctly identify majority of the explanations associated with the driving decisions. This indicates that the model is able to capture the correct causal relationship between the driving decisions and the driving environment, and this capability is useful to enhance the user trust in the automated system.

4.4.5 Potential to identify the limitations of the existing model

From the last two columns in Figure 9, it can be inferred that a weakness of the model is its inability to predict the explanations pertaining to the lane location of the vehicle (the model fails to identify the vehicle is on the left-turning lane in both cases). This problem may be due to the lack of training data associated with this explanation as the original BDD-OIA data set is unbalanced with relatively very few examples indicating that the vehicle should make a turn as it is on the corresponding lane. This can be mitigated by further enriching the dataset by collecting data instances regarding these sparse cases and incrementally training the existing model. Therefore, the proposed model can potentially identify not only its limitations but also the direction of its improvement in a human-understandable manner. This property does generally not exist in most other DL models in the existing literature.

5. Conclusion

In this paper, we propose a novel architecture to generate driving actions as well as explanations based on images, to facilitate autonomous driving. The objective is to mitigate the low interpretability nature of DL-based computer vision models and ultimately, to enhance user trust of autonomous driving systems. The proposed architecture uses the Transformer model (i.e. the MHSA module) to imitate the peripheral vision of humans. The results from the experiments demonstrate that the proposed model outperforms all the baseline models in terms of prediction accuracy and training time.

In the process of addressing these broad objectives, the study evaluated the relative importance of the global features and the local features as well as the appropriate visual attention mechanism for feature engineering. The experiment results suggest that based on the BDD-OIA data set used in the study:

- global features are relatively more important than regional features; and
- the soft attention (Transformer) is superior to hard attention (region selection).

These results are consistent with intuition: for the high-level driving decisions (go straight, slow down/stop, etc.) the peripheral vision (emulated by the global attention) that can achieve long-range correlation and can quickly grasp the overall semantics in the driving environment is found to be more essential compared to foveal vision which specifically focuses on a relatively small region. Therefore, in the development of actual vision-based autonomous driving systems, it is recommended that the designers assign higher priority to the overall information and create the appropriate attention mechanism to enhance the global features.

In the contexts of situational awareness and driver assistance, the proposed model can perform as a driving alarm system for both human-driven vehicles and autonomous vehicles because it is capable of quickly understanding/characterizing the environment and identifying any infeasible driving actions. In addition, the extra explanation head of the proposed model provides an extra channel for sanity checks to guarantee that the model learns the ideal causal relationships. This provision is critical in the development of autonomous systems.

Moving forward to the future work, the proposed model can be further improved by incorporating and fusing other sources (sensor types) such as LiDAR point clouds and information from vehicle-to-vehicle (V2V) connectivity. In this context, the camera is a powerful sensor that can capture a number of semantics in the driving environment, but is vulnerable to occlusion, poor illumination, reflection and so on, and V2V connectivity can address these limitations. V2V provides more straightforward information on the speed, speed change rate and location of neighboring vehicles, and this information can be used directly in the vehicle motion planning module without perception requirements. The fusion of information from multiple sources imparts to the autonomous driving system, the virtues of information redundancy, resilience to possible sensor misfunction and an added layer of system reliability and occupant safety.

References

Alwosheel, A., van Cranenburgh, S. and Chorus, C.G. (2021), "Why did you predict that? Towards explainable artificial neural networks for travel demand analysis", *Transportation Research Part C: Emerging Technologies*, Vol. 128, doi: 10.1016/j.trc.2021.103143.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P. Jackel, L.D. *et al.* (2016), "End to end learning for self-driving cars", pp. 1-9.

Bustos, C., Rhoads, D., Solé-Ribalta, A., Masip, D., Arenas, A., Lapedriza, A. and Borge-Holthoefer, J. (2021), "Explainable, automated urban interventions to improve pedestrian and vehicle safety", *Transportation Research Part C: Emerging Technologies*, Vol. 125, doi: 10.1016/j.trc.2021.103018.

Chen, S., Leng, Y. and Labi, S. (2019), "A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 35 No. 4, doi: 10.1111/mice.12495.

Chen, S., Dong, J., Ha, P., Li, Y. and Labi, S. (2021), "Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 36 No. 7, doi: 10.1111/mice.12702.

Cui, Z., Henrickson, K., Ke, R. and Wang, Y. (2019), "Traffic graph convolutional recurrent neural network: a deep learning framework for network-scale traffic learning and forecasting", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21 No. 11, doi: 10.1109/tits.2019.2950416.

Dong, J., Chen, S., Zong, S., Chen, T. and Labi, S. (2021b), "Image transformer for explainable autonomous driving system", *In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, pp. 2732-2737.

Dong, J., Chen, S., Li, Y., Du, R., Steinfeld, A. and Labi, S. (2021a), "Space-weighted information fusion using deep reinforcement learning: the context of tactical control of lane-changing autonomous vehicles and connectivity range assessment", *Transportation Research Part C: Emerging Technologies*, Vol. 128, doi: 10.1016/j.trc.2021.103192.

Dong, J., Chen, S., Li, Y., Ha, P.Y.J., Du, R., Steinfeld, A. and Labi, S. (2020), "Spatio-weighted information fusion and DRL-based control for connected autonomous vehicles", *2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020*, doi: 10.1109/ITSC45102.2020.9294550.

Doran, D., Schulz, S. and Besold, T.R. (2018), "What does explainable AI really mean? A new conceptualization of perspectives", *CEUR Workshop Proceedings*.

Du, R., Chen, S., Dong, J., Ha, P.Y.J. and Labi, S. (2021), "GAQ-EBkSP: a DRL-based urban traffic dynamic rerouting framework using fog-cloud architecture", doi: 10.1109/isc253183.2021.9562832.

FHWA (2019), "Evaluation methods and techniques: advanced transportation and congestion management technologies deployment program, tech", *Rep. Nr. FHWA-HOP-19-053, Prepared by the Volpe National Transportation Syst*, Washington, DC.

Ha, P., Chen, S., Du, R., Dong, J., Li, Y. and Labi, S. (2020), "Vehicle connectivity and automation: a sibling relationship", *Frontiers in Built Environment*, Vol. 6, doi: 10.3389/fbuil.2020.590036.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2016.90.

Horgan, J., Hughes, C., McDonald, J. and Yogamani, S. (2015), "Vision-based driver assistance systems: survey, taxonomy and advances", *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, doi: 10.1109/ITSC.2015.329.

Hou, R., Jeong, S., Lynch, J.P. and Law, K.H. (2020), "Cyber-physical system architecture for automating the mapping of truck loads to bridge behavior using computer vision in connected highway corridors", *Transportation Research Part C: Emerging Technologies*, Vol. 111, doi: 10.1016/j.trc.2019.11.024.

Hu, H., Zhao, T., Wang, Q., Gao, F. and He, L. (2020), "R-CNN based 3D object detection for autonomous driving", *CICTP 2020: Transportation Evolution Impacting Future Mobility – Selected Papers from the 20th COTA International Conference of Transportation Professionals*, doi: 10.1061/9780784483053.077.

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2018), "Calibrating trust through knowledge: introducing the concept of informed safety for automation in vehicles", *Transportation Research Part C: Emerging Technologies*, Vol. 96, doi: 10.1016/j.trc.2018.07.001.

Kim, J. and Canny, J. (2017), "Interpretable learning for self-driving cars by visualizing causal attention", *Proceedings of the IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.320.

Kim, J., Moon, S., Rohrbach, A., Darrell, T. and Canny, J. (2020), "Advisable learning for self-driving vehicles by internalizing observation-to-action rules", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR42600.2020.00968.

Ku, J., Pon, A.D. and Waslander, S.L. (2019), "Monocular 3D object detection leveraging accurate proposals and shape reconstruction", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2019.01214.

Lioris, J., Pedarsani, R., Tascikaraoglu, F.Y. and Varaiya, P. (2017), "Platoons of connected vehicles can double throughput in urban roads", *Transportation Research Part C: Emerging Technologies*, Vol. 77, doi: 10.1016/j.trc.2017.01.023.

Litman, T. (2014), "*Autonomous Vehicle Implementation Predictions: Implications for Transport Planning*", *Transportation Research Board Annual Meeting*, doi: 10.1613/jair.301.

Liu, Y., Liu, Z. and Jia, R. (2019), "DeepPF: a deep learning based architecture for metro passenger flow prediction", *Transportation Research Part C: Emerging Technologies*, Vol. 101, doi: 10.1016/j.trc.2019.01.027.

McCausland, P. (2019), "Self-driving uber car that hit and killed woman did not recognize that pedestrians jaywalk", *NBC News*, pp. 3-5.

NTSB (2019), "Collision between vehicle controlled by developmental automated driving system and pedestrian", Highway Accident Report NTSB/HAR19/03 Washington, DC.

Peng, B., Keskin, M.F., Kulcsár, B. and Wymeersch, H. (2021), "Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning", *Communications in Transportation Research*, Vol. 1, doi: 10.1016/j.commtr.2021.100017.

Ren, S., He, K., Girshick, R. and Sun, J. (2017), "Faster R-CNN: towards real-time object detection with region proposal networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39 No. 6, doi: 10.1109/TPAMI.2016.2577031.

Rosenholtz, R. (2016), "Capabilities and limitations of peripheral vision", *Annual Review of Vision Science*, Vol. 2 No. 1, doi: 10.1146/annurev-vision-082114-035733.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. (2018), "MobileNetV2: inverted residuals and linear bottlenecks", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2018.00474.

Schwarting, W., Alonso-Mora, J. and Rus, D. (2018), "Planning and decision-making for autonomous vehicles", *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 1 No. 1, pp. 187-210.

Sinha, K.C. and Labi, S. (2007), "Transportation decision making: principles of project evaluation and programming, transportation decision making: principles of project evaluation and programming", doi: 10.1002/9780470168073.

Sowmya Shree, B.V. and Karthikeyan, A. (2018), "Computer vision based advanced driver assistance system algorithms with optimization techniques-a review", *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, doi: 10.1109/ICECA.2018.8474604.

Talpaert, V., Sobh, I., Ravi Kiran, B., Mannion, P., Yogamani, S., El-Sallab, A. and Perez, P. (2019), "Exploring applications of deep reinforcement learning for real-world autonomous driving systems", *VISIGRAPP 2019 – Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, doi: 10.5220/0007520305640572.

TRB (2018), "Socioeconomic impacts of automated and connected vehicle: summary of the sixth EU – US", *Transportation Research Symposium*, *Transportation Research Board Conference Proceedings*.

TRB (2019), "TRB forum on preparing for automated vehicles and shared mobility: mini-workshop on the importance and role of connectivity", *Transportation Research Circular*.

Veres, S.M., Molnar, L., Lincoln, N.K. and Morice, C.P. (2011), "Autonomous vehicle control systems – a review of decision making", *Proceedings of the Institution of Mechanical Engineers. Part I: Journal of Systems and Control Engineering*, doi: 10.1177/2041304110394727.

Wolfe, B., Dobres, J., Rosenholtz, R. and Reimer, B. (2017), "More than the useful field: considering peripheral vision in driving", *Applied Ergonomics*, Vol. 65, doi: 10.1016/j.apergo.2017.07.009.

World Bank (2005), "A framework for the economic evaluation of transport projects, transport notes".

Xia, Y., Kim, J., Canny, J., Zipser, K., Canas-Bajo, T. and Whitney, D. (2020), "Periphery-fovea multi-resolution driving model guided by human attention", *Proceedings – 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, doi: 10.1109/WACV45572.2020.9093524.

Xing, Y., Lv, C., Cao, D. and Velenis, E. (2021), "Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles", *Transportation Research Part C: Emerging Technologies*, Vol. 130, doi: 10.1016/j.trc.2021.103288.

Xu, H., Gao, Y., Yu, F. and Darrell, T. (2017), "End-to-end learning of driving models from large-scale video datasets", *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, doi: 10.1109/CVPR.2017.376.

Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y. and Vasconcelos, N. (2020), "Explainable object-induced action decision for autonomous vehicles", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR42600.2020.00954.

Yadron, D. and Tynan, D. (2016), "Tesla driver dies in first fatal crash while using autopilot mode", *The Guardian*.

Yu, B., Lee, Y. and Sohn, K. (2020a), "Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN)", *Transportation Research Part C: Emerging Technologies*, Vol. 114, doi: 10.1016/j.trc.2020.02.013.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V. and Darrell, T. (2020b), "BDD100K: a diverse driving dataset for heterogeneous multitask learning", *Proceedings of the IEEE Computer Society Conference* *on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR42600.2020.00271.

Zhao, H., Jia, J. and Koltun, V. (2020), "Exploring self-attention for image recognition", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR42600.2020.01009.

Zhou, F., Li, L., Zhang, K. and Trajcevski, G. (2021), "Urban flow prediction with spatial–temporal neural ODEs", *Transportation Research Part C: Emerging Technologies*, Vol. 124, doi: 10.1016/j.trc.2020.102912.

Zhu, W., Wu, J., Fu, T., Wang, J., Zhang, J. and Shangguan, Q. (2021), "Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP", *Journal of Intelligent and Connected Vehicles*, Vol. 4 No. 2, doi: 10.1108/jicv-03-2021-0004.

Zhuang, L., Wang, L., Zhang, Z. and Tsui, K.L. (2018), "Automated vision inspection of rail surface cracks: a double-layer data-driven framework", *Transportation Research Part C: Emerging Technologies*, Vol. 92, doi: 10.1016/j.trc.2018.05.007.

Corresponding author

Sikai Chen can be contacted at: chen1670@purdue.edu