IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 March 2024, accepted 8 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3400589*

RESEARCH ARTICLE

# DRUformer: Enhancing Driving Scene Important Object Detection With Driving Scene Relationship Understanding

**YINGJIE NIU[1], (Student Member, IEEE), MING DING[1], (Member, IEEE), KEISUKE FUJII[1], (Member, IEEE), KENTO OHTANI[1], ALEXANDER CARBALLO[1,2], (Member, IEEE), AND KAZUYA TAKEDA[1,3], (Senior Member, IEEE)**

[1]Graduate School of Informatics, Nagoya University, Nagoya 464-8603, Japan
[2]Graduate School of Engineering, Gifu University, Gifu 501-1112, Japan
[3]Tier IV Inc., Tokyo 140-0001, Japan

Corresponding author: Yingjie Niu (niu.yingjie@g.sp.m.is.nagoya-u.ac.jp)

This work was supported in part by Nagoya University, in part by Japan Society for the Promotion of Science (JSPS), in part by Japan Science and Technology Agency (JST) under Grant JPMJFS2120, and in part by JSPS KAKENHI under Grant JP21H04892 and Grant JP21K12073.

**ABSTRACT** Traffic accidents frequently lead to fatal injuries, claiming millions of lives every year. To mitigate driving hazards and ensure personal safety, it is crucial to assist vehicles in anticipating the objects in the traffic scene (treated here as important objects) which may pose a threat during the driving task. Previous research on important object detection primarily assessed the importance of individual participants, treating them as independent entities and frequently neglecting the interconnections among these participants. Unfortunately, this approach has proven less effective in detecting important objects in complex scenarios. In this work, we introduce Driving scene Relationship Understanding transformer (DRUformer), designed to enhance the important object detection task. The DRUformer is a transformer-based multi-modal important object detection model that takes into account the relationships between all the participants in the driving scenario. Recognizing that driving intention also significantly affects the detection of important objects during driving, we have incorporated a module for embedding driving intention. To assess the performance of our approach, we conducted comparative experiments on the DRAMA dataset, comparing our model against other state-of-the-art (SOTA) models. The results demonstrated a noteworthy 16.2% improvement in mIoU and a substantial 12.3% boost in ACC compared to SOTA methods. Furthermore, we conducted a qualitative analysis of our model's ability to detect important objects across different road scenarios and classes, highlighting its effectiveness in diverse contexts. Finally, we conducted various ablation studies to assess the efficiency of the proposed modules in our DRUformer model. Through extensive experimentation, it has been demonstrated that our model performs exceptionally well in the task of driving scene important object localization. The code is publicly available on the following link: https://github.com/oniu-uin0/DRUformer

**INDEX TERMS** Important object detection, multi-modal, driving scene relationship understanding, driving intention.

## I. INTRODUCTION

Recent years have seen notable advancements in the evolution of Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicles (AVs). The overarching goal of this research is to establish a service system that prioritizes safety and comfort for humanity [1]. However, there remains room for enhancement in ensuring driving safety. From 2014 to September 20, 2023, the DMV received 655 Autonomous Vehicle Collision Reports [2]. According to the World Health

The associate editor coordinating the review of this manuscript and approving it for publication was Shunfeng Cheng.


(a) (b)

**FIGURE 1.** Overview of our work. As shown in the (a), which is easily understandable for human drivers, the significance of the first pedestrian outweighs that of the second, as humans recognize that pedestrian 1 is walking within the driving lane, whereas pedestrian 2 is not. Human drivers grasp an inherent relationship between pedestrians and line markings. As depicted in (b), even within the same intersection, the significance of objects can vary based on fluctuations in the driver's intention: if the driver intends to turn left, pedestrian 3 becomes the most important, whereas if the intention is to go straight, pedestrian 4 takes precedence.

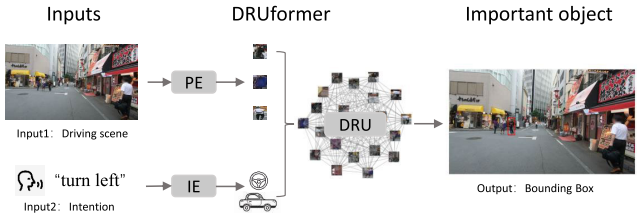


**FIGURE 2.** Overview of our pipeline. Our DRUformer model takes both the "driving scene" and the driving "intention" as input, to ascertain the most "important object" within the scene while considering the relationships among all the participants.

Organization (WHO), the annual global death toll due to road accidents stands at a staggering 1.3 million. More than half of all road traffic deaths and injuries involve vulnerable road users, such as pedestrians, cyclists, motorcyclists, and their passengers [3], [4]. Therefore, developing a system to predict important objects in driving scenarios is crucial to advance ADAS and AV technology [5].

Driving scene Important Object localization (IOL) [6] involves assessing all participants in a driving scenario to predict objects that could significantly influence future control decisions. The potential applications of IOL are wide-ranging. It can be seamlessly integrated into ADAS to help drivers identify hazardous objects and mitigate potential dangers. Additionally, IOL systems can support AVs in detecting potential hazards, streamlining subsequent tasks like trajectory prediction, decision-making, and motion planning.

Currently, a considerable body of research [7], [8], [9], [10], [11], [12], [13] simplifies IOL as a mere object detection task. This approach treats all participants as independent entities and directly predicts the importance of each object to the ego-vehicle. They frequently neglect the intricate relationships among all participants, whereas human drivers instinctively consider these interrelations when predicting the importance of objects. This kind of methods fails to distinguish the importance difference illustrated in Fig. 1. As shown in fig. 1(a), which is easily understandable for human drivers, the significance of the first pedestrian outweighs that of the second, as humans recognize that pedestrian 1 is walking within the driving lane, whereas pedestrian 2 is not. Human drivers grasp an inherent relationship between pedestrians and line markings. In actual driving situations, human drivers take into account not only the influence of individual objects on the ego-vehicle but also the interactions between these objects.

Consequently, comprehending the interactions among participants in driving scenes holds paramount importance for IOL. There are existing methods [13], [14], [15], [16], [17], [18], [19] designed to instruct models about relationships within scenes. However, a common limitation among these methods is their reliance on manually annotated relationship labels to guide machine learning models. Unfortunately, manual annotation falls short in capturing all object relationships. Some methods focus solely on annotating spatial relationships within a single scene, while others concentrate on the relationships of specific participants, neglecting many crucial participants. Moreover, manually defined relationships are crafted for human comprehension but may not necessarily be optimal for machine understanding. Hence, the introduction of a "driving scene relationship understanding" module becomes crucial for enabling the model to autonomously learn interrelationships among all objects.

While previous research has primarily focused on enhancing the extraction of visual information from driving scenes to improve IOL accuracy, it's important to acknowledge that the significance of key objects can shift in real driving situations, influenced by the driver's intentions. As depicted in Fig. 1(b), even within the same intersection, the significance of objects can vary based on fluctuations in the driver's intention. If the driver intends to turn left, pedestrian 3 becomes the most important, whereas if the intention is to go straight, pedestrian 4 takes precedence. Hence, predicting important objects should consider not only the driving scene but also the driver's intention.

Another critical concern is that current IOL models heavily depend on CNN architectures. However, for AV tasks, Hu et al. [20] have illustrated that varying different modal information across tasks may result in low information exchange efficiency during CNN feature extraction. Hence, it is essential to propose a IOL model based on an architecture that distinguishes itself from CNNs. Within this architecture, the IOL model should adeptly blend information from both the driving scene and driving intention. Moreover, it must seamlessly transmit this combined information to the subsequent module for relationship understanding without any data loss.

To address the aforementioned issues, we introduce a model named **D**riving scene **R**elationship **U**nderstanding trans**former** (DRUformer) for the IOL task, as depicted in Fig. 2. Our model is designed to simultaneously integrate both the driving scene (driving frame) and driving intention (textual intention command) while taking into account the relationships between participants, ultimately enabling the prediction of important objects within the driving scene.

Our model primarily comprises three modules: the Participants Extractor Module (PE), which extracts information about all participants in the driving scene; the Intention Extractor Module (IE), responsible for extracting driving commands related to the driver's intention; and the Driving scene Relationship Understanding Module (DRU), which focuses on learning the interrelations between all objects without the mannual relationship definition. Transformer-based algorithms have been extensively proven as the most suitable framework for multi-modal tasks in recent studies [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. Therefore the entire approach is constructed upon the Transformer framework, facilitating information exchange optimization between all modules and enhancing downstream tasks. To evaluate the effectiveness of our proposed model, we conducted comprehensive qualitative and quantitative analyses on the largest existing important object detection dataset, the DRAMA [6] dataset.

In summary, our contributions are as follows:

- We propose a novel model that predicts important objects by taking into account the relationships between all driving participants as well as the driver's driving intention.
- We introduce a driving scene relationship understanding module to learning the relationships between the participants without the need for manual relationship definition.
- We conducted extensive quantitative and qualitative analyses on the DRAMA dataset, and our model significantly outperforms the baseline, resulting in substantial improvements in the IOL task.
- To facilitate reproducibility and further research, we contribute our code to the community.

The subsequent chapters will be organized as follows: In section II, we will provide an overview of current developments related to the IOL task and research related to relationship understanding. Section III will outline our DRUformer model and the construction of its individual modules. In section IV, we will detail the datasets employed in our experiments, the experimental design, and the results achieved. In section V, we will discuss the limitations and future directions for our work. Section VI will offer a summary of the contributions made throughout our paper.

## II. RELATED WORK

### A. RISK OBJECT DETECTION

The task of risk object detection can primarily be categorized into two approaches: direct detection and indirect detection.

Direct detection aims to use the powerful regression capabilities of neural networks to mimic human-annotated datasets of important objects based on supervised tasks. In essence, these methods classify all objects into two categories: important and non-important. Several studies [7], [8], [9], [10], [31] focus on utilizing human gaze (attention) information from experienced drivers as the supervised training labels. This label information is used to predict pixel-level attention regions through neural networks. However, this approach encounters two significant challenges. Firstly, human drivers may become distracted during driving, leading to their gaze fixating on objects unrelated to the driving task, such as interesting billboards. Secondly, the pixel-level information may not always effectively clustered around objects. Therefore, Zhang et al. [13], Karim et al. [12], and Malla et al. [6], [11] have adopted the use of object-level bounding box information as the learning target. This shift in focus enhances the prediction of risk objects, ensuring that it primarily concentrates on objects relevant to driving.

Indirect detection methods predict risk objects or risk regions through proxy tasks such as brake prediction and steering wheel angle prediction. Zhang et al. [13], Wang et al. [32] and Li et al. [33] involves pre-training a model for driving behavior prediction and subsequently identifying hazardous objects by removing extraneous objects from the scene. The truly risky object is believed to have a significant impact on driving outputs. Kim et al. [34], [35] focus on training an end-to-end model for image-to-steering angle prediction and conducts causality analysis using attention maps to forecast hazardous pixel areas within the CNN model. It's worth noting that indirect prediction methods are primarily intended for offline use and are particularly well-suited for explainability AI but not optimal for online IOL tasks.

While the previously mentioned methods have undeniably advanced the field of IOL, they still grapple with the following issues:

- They often overlook the intricate relationships among all participants, whereas human drivers instinctively consider these interrelations when predicting the importance of objects.
- These methods do not take into account the importance of human driving intentions, especially in advanced autonomous vehicles.
- Most of their models rely on CNN architectures, which may not be ideal for conducting multi-modal research and facilitating information exchange for downstream tasks.

To address these challenges, we propose a transformer-based multi-modal model that takes into account both the interrelationships between the participants as well as the driver's driving intention.

### B. RELATIONSHIP UNDERSTANDING

The Relationship Understanding (RU) task involves comprehending the relationships between objects within a scene through various methodologies. Currently, RU plays a pivotal role in tasks such as Human-Object Interaction (HOI) and Scene-Graph Generation (SG) [13], [14], [15], [16], [17], [18], [19], [36], [37], [38], [39], [40]. RU can be broadly classified into two approaches: two-stage (sequential) and one-stage (parallel) methodologies.

Two-stage methods typically encompass entity detection and relationship classification as two sequential steps. In the initial stage, an off-the-shelf object detector is employed to identify all objects, while in the subsequent stage, the detected objects are paired, and relationships between these paired objects are predicted. These frameworks aim to optimize object detection separately, either to enhance object detection accuracy or to improve HOI tasks through interaction classification optimization. Nonetheless, the primary challenge with this pipeline lies in the fact that the two stages cannot be concurrently fine-tuned, often resulting in decreased detection accuracy and efficiency.

Conversely, one-stage methods treat HOI as a Set Prediction problem, simultaneously predicting object pairs boxes and relationships between objects. This approach offers a more direct and efficient methodology with reduced time complexity [41]. However, one-stage methods frequently require intricate post-processing steps.

In the realm of autonomous driving, Yu et al. [16] have employed a two-stage SG method to define spatial location relationships in highway driving situations, transforming the driving scene into a driving scene-graph. This approach is limited to specific scenarios and only addresses spatial location relationships (e.g., left, front, right). Another study by Tian et al. [15] also employs a two-stage SG method for the driving SG task, encompassing a broader range of scenarios, including pedestrians, bicycles, and more agents. However, this method still falls short in accounting for the intricate relationships between the driving vehicle, traffic signals, and driving direction.

Despite the significant advancements brought about by the aforementioned methods in the DRU task, they still confront the following challenges:

- The relationships defined within their scenes often rely on manual definitions, which prove insufficient for the complexity of the driving environment, where not all relationships can be predefined.
- The definitions of relationships between objects predominantly focus on spatial relationships, whereas relationships can be more multifaceted. For instance, the connection between a pedestrian and a traffic signal exists even if they are spatially distant.
- Existing driving RU methods primarily adhere to a two-stage approach, resulting in relationships that are often constrained by the outcomes of object detection. Simultaneously optimizing relationship generation and object detection remains a challenge in this context.

To tackle these challenges, we introduce a DRU module that empowers our DRUformer to learn relationships between all participants without relying on human-defined labels. To concurrently enhance participant detection and relationship comprehension, we treat each participant as a transformer token, allowing for direct integration into the DRU module.

## III. METHOD

To leverage the success of the transformer-based multimodal approach, we present a transformer-based multimodal model designed for the detection of critical objects in driving scenes, which we have named *DRUformer*. The model's architecture is depicted in Fig. 3. The DRUformer model comprises three key modules:

- The *Participants Extractor* module (PE) is tasked with capturing both the positional and semantic information of objects that are relevant to the act of driving.
- The *Intention Extractor* module (IE) is specially designed to collect the driving intention from the driver intention command.
- The *Driving relationship understanding* module (DRU) is dedicated to comprehending the interactions that transpire among all participants within the driving scene without the need for manual definition.

### A. PARTICIPANTS EXTRACTOR MODULE

Before constructing relationships between all objects, it is necessary to first detect all participants present in the scene. This module is specifically designed to extract the participants in the driving scene.

The PE module is designed based on the powerful transformer-based object detection model called DETR [42]. PE comprises a standard CNN backbone referred to as $f_c$, a standard transformer encoder denoted as $f_e$, a standard transformer decoder designated as $f_d$, and a simple feed forward network (FFN) to make the participant bounding box prediction. We should note that the FFN part is only utilized in the pre-training step and does not participate in the main training step.

For the driving scene $\boldsymbol{I}$ with size $H \times W \times 3$, we utilize $f_c$ to extract a lower-resolution feature map $\boldsymbol{z} \in \mathbb{R}^{H_S \times W_S \times D_S}$ encompassing the entire scene, $H_S \times W_S$ represents the size of the feature map and $D_S$ represents the dimension. In this phase, $\boldsymbol{I}$ is downsampled into global spatial features with dimensions $H_S \times W_S$, typical $H_S, W_S = \frac{H}{32}\frac{W}{32}$. The encoder $f_e$ expects a sequence as input, hence we flatten and embedding the spatial feature $\boldsymbol{z}$ into one dimension features $\boldsymbol{F} = \left\{\boldsymbol{f_i} \,|\, \boldsymbol{f_i} \in \mathbb{R}^{D_d}\right\}_{i=1}^{H_S \times W_S}$, $D_d$ represents the channel dimension of the features, $D_d$ is often much smaller than $D_S$. These features in $\boldsymbol{F}$ are then passed through $f_e$ for global feature extraction.

In the $f_e$ module, each $f_e$ layer consists of a multi-head self-attention (MHSA) module and a FFN. Since the transformer architecture is permutation-invariant, we integrate positional information from $\boldsymbol{F}$ by inputting the positional encoding $\boldsymbol{p} \in \mathbb{R}^{H_S \times W_S \times D_S}$ into the $f_e$ module. This results in global features $\boldsymbol{S} = \left\{\boldsymbol{m_i} \,|\, \boldsymbol{m_i} \in \mathbb{R}^{D_d}\right\}_{i=1}^{H_S \times W_S}$ that include position information. These features are then used in subsequent decoder stages.

In the case of $f_d$, each $f_d$ layer includes a self-attention module, a cross-attention module, and an FFN. We initialize $N$ queries denoted as $\boldsymbol{Q} = \left\{\boldsymbol{q_i} | \boldsymbol{q_i} \in \mathbb{R}^{D_d}\right\}_{i=1}^{N}$, to extract

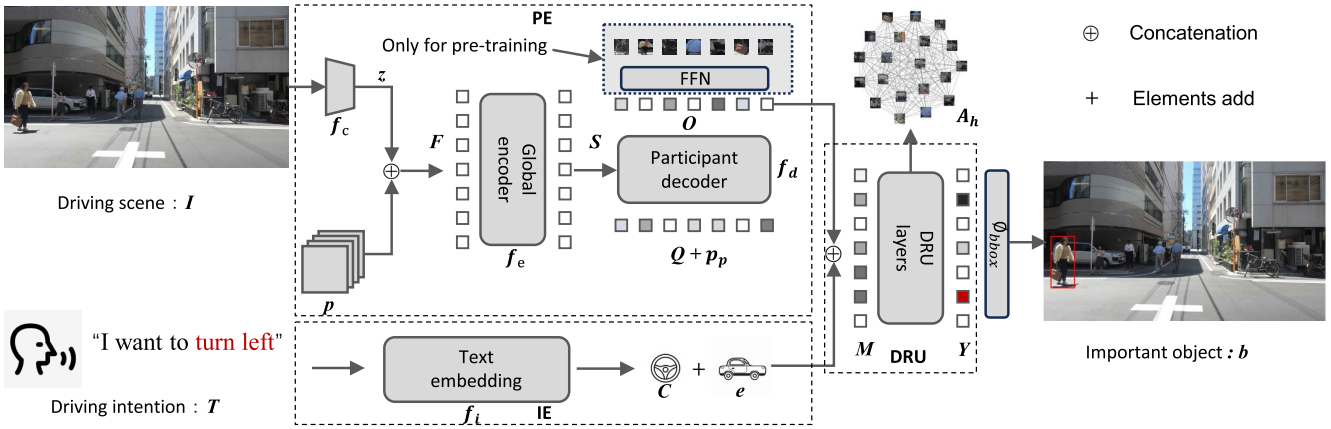


**FIGURE 3.** Architecture of the DRUformer. This is the pipeline of our DRUformer model. We feed the driving scene $I$ and the driving intention $T$ into our DRUformer and obtain the important object $b$. In this context, the "PE" module signifies the participants extractor, the "IE" module signifies the driving intention extractor, and the "DRU" module represents the driving relationship understanding module.

object information from the global features $\boldsymbol{S}$. Since the decoder is also permutation-invariant, to incorporate positional information into these $\boldsymbol{Q}$ features, we include positional encoding $\boldsymbol{p}_p \in \mathbb{R}^{H_S \times W_S \times D_S}$. After processing through the $f_d$ module, we obtain $N$ scene participants tokens, denoted as $\boldsymbol{O} = \{\boldsymbol{o_i} | \boldsymbol{o_i} \in \mathbb{R}^{D_d}\}_{i=1}^{N}$, which includes both the positional and semantic information of $N$ participants in the scene. The $N$ scene participants tokens $\boldsymbol{O}$ are then independently decoded into box coordinates and class labels by a FFN, resulting $N$ predictions. As mentioned earlier, we aim to simultaneously optimize relationship generation and object detection. Therefore, we exclusively employ the FFN only during the pre-training phase. In the model training stage, we rely solely on scene participants tokens $\boldsymbol{O}$ for information propagation, omitting the use of FFN-generated boxes and class information.

The PE is calculated as follows:

$$\boldsymbol{F} = \text{flatten}(f_c(\boldsymbol{I})), \tag{1}$$

$$\boldsymbol{S} = f_e(\boldsymbol{F} + \boldsymbol{p}), \tag{2}$$

$$\boldsymbol{O} = f_d(\boldsymbol{Q} + \boldsymbol{p}_p, \boldsymbol{S}). \tag{3}$$

where flatten represents the flatten operation.

### B. INTENTION EXTRACTOR MODULE

As mentioned earlier, driving direction is also crucial for the task of important object detection. This module is designed to extract the driver's driving intention command.

In our research, the "Intention Extractor" (IE) module, denoted as $f_i$, is constructed based on a text embedding model. The driver's text intention command, labeled in the DRAMA dataset and represented as $\boldsymbol{T}$, is tokenized by the IE module into an intention command token, denoted as $C \in \mathbb{R}^{D_d}$. It's important to note that, for the convenience of subsequent relationship understanding part, $\boldsymbol{C}$ and $\boldsymbol{o_i}$ should have the same dimensionality. The calculation formula is as follows:

$$\boldsymbol{C} = f_i(\boldsymbol{T}). \tag{4}$$

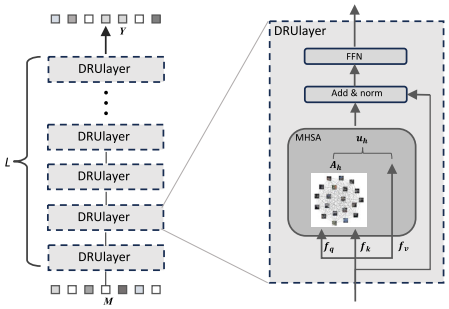


**FIGURE 4.** Architecture of our DRU module. $M$ represents all the entities, $Y$ represents all the entities with understanding the interrelationships. MHSA represents the multi-head self-attention module, and FFN represents the feed-forward network.

### C. DRIVING RELATIONSHIP UNDERSTANDING MODULE

To model dense interrelationships between the ego-vehicle and other participants, we employ a self-attention module to learn the mutual relationships between all objects.

Initially, we randomly initialize a learnable token to represent the ego-vehicle token, denoted as $\boldsymbol{e} \in \mathbb{R}^{D_d}$, which serves as a carrier for the intention token $\boldsymbol{C}$, making it easier to learn. Since we aim to learn the relationships among all objects in the scene, we then combine $\boldsymbol{C}$ and $\boldsymbol{e}$ together and concatenate them with the participants $\boldsymbol{O}$ to get all entities $\boldsymbol{M} \in \mathbb{R}^{k \times D_d}$, where $k = N + 1$. This concatenation is expressed as:

$$\boldsymbol{M} = \text{norm}\,(\boldsymbol{C} + \boldsymbol{e}) \oplus \boldsymbol{O}. \tag{5}$$

where norm $(\cdot)$ represents the layer normalization.

Subsequently, we feed all the entities in $\boldsymbol{M}$ into our DRU module to learn their relationships with each other. Our DRU module consists of $L$ layers of DRU layers. Figure 4 visually illustrates the architecture of our DRU module. Each DRU layer comprises MHSA and a FFN, with $H$ heads in the MHSA. Each relationship head deals with the subspace $\boldsymbol{m} \in \mathbb{R}^{k \times D_h}$, where $D_h = \frac{D_d}{H}$. The understanding mechanism is

calculated as follows:

$$\boldsymbol{A}_h = \text{softmax}\left(\frac{f_q(\boldsymbol{m}) f_k(\boldsymbol{m})^T}{\sqrt{D_h}}\right), \tag{6}$$

$$\boldsymbol{u}_h = \boldsymbol{A}_h \cdot f_v(\boldsymbol{m}). \tag{7}$$

where $f_q, f_k$, and $f_v$ represent the linear transformation layers, and $\boldsymbol{A}_h \in \mathbb{R}^{k\times k}$ is the entity relationship map for each heads. It should be noted that in our research, the strength of the relationship between two objects is represented as the cosine similarity between the two objects, with a higher similarity indicating a stronger relationship. Each head's output $\boldsymbol{u}_h \in \mathbb{R}^{k\times D_h}$ is already understanding the relationship.

Finally, the output of each relationship head needs to be concatenated together and passed through an FFN to obtain the output of the DRU layer, denoted as $\boldsymbol{Y} \in \mathbb{R}^{k\times D_d}$, at which point $\boldsymbol{Y}$ has comprehended the relationships between all objects.

At this stage, the bounding boxes of the important objects $\boldsymbol{b} \in \mathbb{R}^{k\times 4}$ in the scene can be predicted using a predictor FFN $\phi_{bbox}(\cdot)$, as shown in the following formula:

$$\boldsymbol{b} = \phi_{bbox}(Y). \tag{8}$$

### D. TRAINING

Following the set-based training process of DETR, we first match each ground truth important object bounding box with its best-matching prediction by the bipartite matching with the Hungarian matching algorithm [42]. Then the loss is produced between the matched predictions and the corresponding ground truths for the final back-propagation.

Similar to DETR, the loss of DRUformer is composed of 3 parts: the box regression loss $\mathcal{L}_b$, the intersection-over-union loss $\mathcal{L}_{GIoU}$ [43], and the object important class loss $\mathcal{L}_c$. Loss function is shown as follows:

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_{GIoU} \mathcal{L}_{GIoU} + \lambda_c \mathcal{L}_c. \tag{9}$$

where $\lambda_b$, $\lambda_{GIoU}$ and $\lambda_c$ are the hyper-parameters for adjusting the weights of each loss.

## IV. EXPERIMENT

The code of our paper and more detection results is available here: https://github.com/oniu-uin0/DRUformer.

### A. EXPERIMENT SETTING

#### 1) DATASET

To accurately evaluate the capabilities of our proposed model, we chose to conduct tests on the largest existing important object evaluation dataset, DRAMA. This dataset was collected by HONDA Corporation in the Tokyo region of Japan and comprises a total of 17,785 scenarios, encompassing various road environments such as wider roads, intersections, and narrow streets. This diversity allows us to evaluate our model under different road conditions. The majority of scenarios are labeled with the driving direction of the ego-vehicle in that scenario and the bounding box information of important object. The driving direction serves as the driving intention of our model. The important objects in the dataset include vehicles, pedestrians or cyclists, and infrastructure, offering a diverse range of object sizes for testing.

To ensure a robust evaluation, we shuffled the dataset and then divided it into training, validation, and test sets in a 70:15:15 ratio for training and testing purposes.

#### 2) EVALUATION METRICS

In our study, we employed three standard evaluation metrics, Frames Per Second (FPS), Accuracy (ACC) and mean Intersection over Union (mIoU), to better evaluate the overlap between the predicted bounding boxes of important object and the manually annotated important object by experts.

mIoU assesses the mean overlap between the predicted bounding boxes, denoted as $y_{pred}$, and the ground truth bounding boxes, represented as $y_{label}$. A higher mIoU value means more precise predictions of the bounding boxes. The calculation formula for mIoU is as follows:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{I(y_{pred}, y_{label})}{U(y_{pred}, y_{label})} \tag{10}$$

where $N$ represents the number of samples, and $I(y_{pred}, y_{label})$ and $U(y_{pred}, y_{label})$ represents the Intersection and Union between $y_{pred}$ and $y_{label}$ for each sample, respectively.

ACC evaluates whether our model correctly identifies the presence of important objects in the scene and accurately deter mines their locations within the image. A higher ACC indicates a more effective ability to predice the presence of important objects. The calculation formula for ACC is as follows:

$$ACC = \frac{1}{N}\sum_{i=1}^{N}[y_{pred} = y_{label}], \tag{11}$$

where N represents the number of samples, and $[y_{pred} = y_{label}]$ equals 1 if the $\frac{I(y_{pred}, y_{label})}{U(y_{pred}, y_{label})} > 0.5$.

FPS indicates the inference speed at which the model can process and analyze driving scene frames. Higher FPS values generally denote superior real-time performance and faster processing capabilities.

#### 3) IMPLEMENTATION DETAILS

All of our experiments were conducted on a server with 8 NVIDIA RTX A6000 GPUs. We believe that aligning with the settings of the DETR model provides a strong foundation, as it is a well-established and widely adopted framework in the field of object detection. During the experiment process, the hyperparameters were set as follows: for the CNN part, we utilized the standard ResNet50 architecture and a 6-layer standard Transformer encoder to extract global features. In the decoder part, we used a 6-layer architecture to extract participants in driving scenes. In the DRU module's relationship attention section, we employed an MHSA with 8 heads for relationship understanding. The initial learning

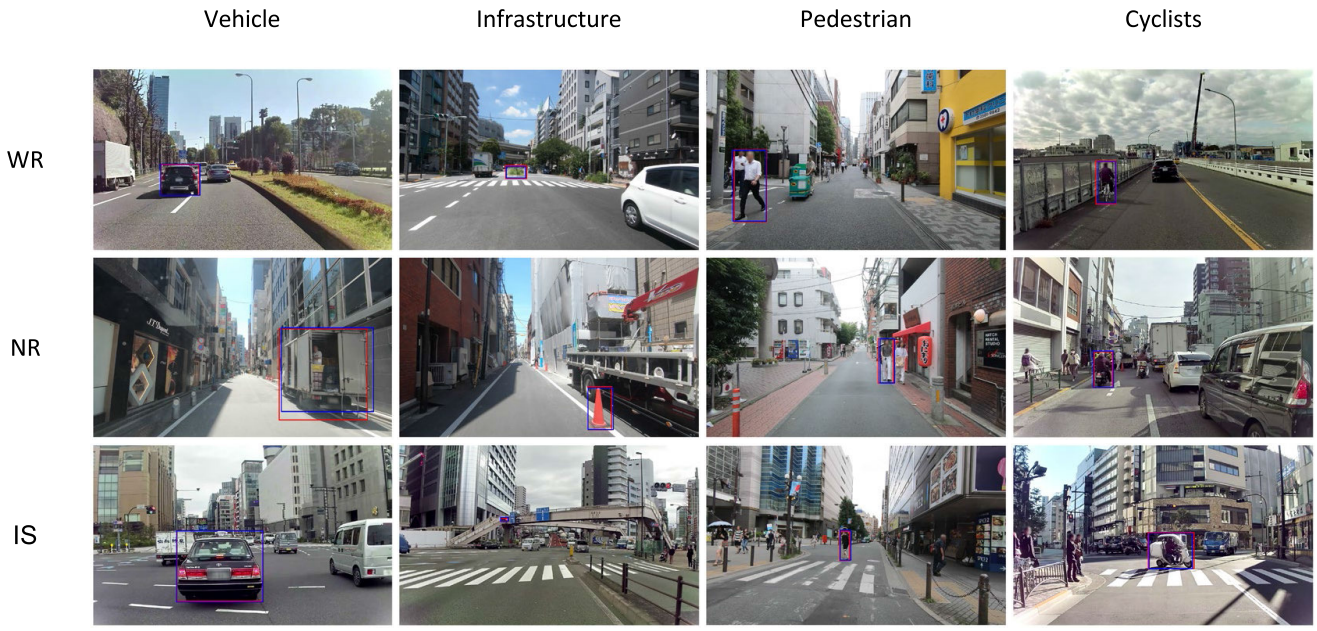


**FIGURE 5.** Qualitative analysis of our DRUformer. These are our model's important object detection results for various scenes and different types of objects. WR represents the Wider Road scene, NR represents the Narrow Road scene, and IS represents the intersection scene. The blue boxes represent the detection results from our model, while the red boxes represent the labels in the dataset.

rate was set to 0.0001, and the weight decaying rate was $10^{-4}$ per 200 epochs. We applied scale augmentation similar to DETR, resizing the input images such that the shortest side was at least 480 pixels and at most 800 pixels, while the longest side was at most 1333 pixels. The size of $L$ was set to 3 layers. During the training phase, the batch size for each epoch was set to 8, and we trained for 400 epochs to obtain our experimental weights.

### B. QUANTITATIVE ANALYSIS

Table 1 presents the localization results of our method and other SOTA baselines on the DRAMA dataset for important object detection. In this table, *ICL* represents Independent Captioning and Localization, *LCP* represents Localization with Captioning Prior, *OF* represents the Optical Flow for the corresponding frames. It is evident that our method, when provided with only the driving frames as input, exhibits a substantial improvement of 12.3% in ACC and 16.2% in mIoU compared to LCP with optical flow. Notably, LCP not only incorporates optical flow but also leverages scene captions for detection assistance, yet it still falls significantly behind our DRU-equipped approach.

In comparison to LCP without optical flow assistance, our method achieves a 14% increase in ACC and a 17.9% boost in mIoU. We also compared our method to SOTA approaches for general object detection tasks such as DINO [44], CO-DETR [45] and DETR [42], and our method exhibited significant improvements over both of these methods. Additionally, we compared the inference speed of our model with these methods. Although our algorithm's inference speed may not be outstanding, it is comparable to these SOTA methods. However, it is important to note that at present, our algorithm is not yet capable of achieving real-time detection. Therefore, further optimization in terms of speed will be necessary in the future. This convincingly demonstrates that our DRUformer delivers outstanding performance in the IOL task on the DRAMA dataset, even without optical flow, showcasing its superior capabilities.

**TABLE 1.** IOL performances comparing with the SOTA methods on the DRAMA dataset.

| Method | mIoU | ACC ($IoU > 0.5$) | FPS |
|---|---|---|---|
| ICL without OF [6] | 0.553 | 0.617 | - |
| ICL [6] | 0.533 | 0.593 | - |
| LCP without OF [6] | 0.597 | 0.667 | - |
| LCP [6] | 0.614 | 0.684 | - |
| CO-DETR [45] | 0.665 | 0.682 | 15 |
| DINO [44] | 0.658 | 0.679 | 24 |
| DETR [42] | 0.644 | 0.671 | 29 |
| DRUformer | **0.776(16.2**% ↑) | **0.807(12.3**% ↑) | 21 |

### C. QUALITATIVE ANALYSIS

Figure 5 presents the position results of our algorithm in various road scenarios, including Wide Road (WR), Narrow Road (NR), and Intersection (IS), with important objects such as vehicles, pedestrians, cyclists, and infrastructure. This visualization allows us to intuitively observe that our method, when predicting important objects, considers not only positional relationships but also the interactions between objects. For example, in the NR scenario pedestrian column, there are two pedestrians walking side by side on the right side. Despite they close to each other, our model understands that only the pedestrian inside the blue box is within the lane (considering the relationship between the pedestrian


**FIGURE 6.** Visual ablation results for driving intention. IN represents the intention. The first column shows the IOL results of our method when excluding the driving intention module, while the second column displays the detection results of our model including the driving intention.

**TABLE 2.** Ablation results for the impact of the driving intention.

| Method | mIoU | ACC ($IoU > 0.5$) |
|---|---|---|
| DRUformer without IN | 0.742 | 0.771 |
| DRUformer with IN | **0.776(3.4% ↑)** | **0.807(3.6% ↑)** |

and the lane). Therefore, the pedestrian inside the blue box is deemed the most important object, rather than both of the pedestrians. Furthermore, our method exhibits excellent detection performance for objects of different sizes. For instance, in the IS scenario, vehicles are larger objects, while infrastructure objects like traffic signals are smaller. Figure 5 also illustrates that our method can predict not only dangerous dynamic objects but also static traffic signals. It is evident that our approach performs well in different driving scenarios and different important objects, providing robust object detection results.

### D. ABLATION STUDIES

We conducted three ablation experiments to validate the effectiveness of the proposed modules.

#### 1) WITH INTENTION VS. WITHOUT INTENTION

Table 2 presents the localization results of our DRUformer in two scenarios: with the driving Intention (IN) and without IN. It is clear that when considering IN, our model shows a significant improvement in the IOL task, with a 3.4% increase in ACC and a 3.6% increase in mIoU compared to the scenario without IN. Even without IN, our method exhibits a 9.8% increase in mIoU and an 8.7% increase in ACC over the SOTA methods. This demonstrates the effectiveness of the Driving Intention.

Figure 6 provides a more intuitive demonstration of the impact of driving intention on our model. It can be observed that without the IN, our model may predict important objects in different directions. For example, in the scenario depicted in Image 3 (a), the model without IN predicts that the two individuals directly in front of the vehicle are important objects, while the actual label indicates the person in the red box on the left is important. However, in this case, it doesn't necessarily imply that the model's prediction is incorrect. If our vehicle needs to move directly forward, then the objects directly in front are indeed the most important. After providing the IN for a left turn, the model can correctly predict the important objects. It is evident that IN is helpful for detecting important objects in scenarios where there is direction ambiguity.

**TABLE 3.** Ablation results for the impact of the DRU.

| Method | mIoU | ACC ($IoU > 0.5$) |
|---|---|---|
| DRUformer without DRU | 0.659 | 0.682 |
| DRUformer | **0.776(11.7% ↑)** | **0.807(12.5% ↑)** |

**TABLE 4.** DRUformer performance with different DRU layers.

| DRU layers | mIoU | ACC ($IoU > 0.5$) |
|---|---|---|
| 1 | 0.694 | 0.717 |
| 2 | 0.712 | 0.749 |
| 3 | **0.776** | **0.807** |
| 4 | 0.748 | 0.776 |
| 5 | 0.733 | 0.758 |
| 6 | 0.686 | 0.745 |

#### 2) WITH DRU MODULE VS. WITHOUT DRU MODULE

Table 3 presents the localization results of our DRUformer model in two scenarios: with the DRU module and without it. It is evident that when considering inter-object relationships, our model demonstrates a significant improvement in the IOL task compared to the scenario without the DRU module. Specifically, there is a 12.5% increase in ACC and an 11.7% increase in mIoU, providing clear evidence of the effectiveness of our DRU module for the IOL task. Even without DRU, our method achieves slightly higher mIoU and similar ACC when compared to the best in SOTA.

Figure 7 visually illustrates the differences between the relationship network learned by our DRU module and other relationship networks based solely on manually defined relationships. We selected ten objects from all the objects detected by the PE module and combined them with the ego-vehicle to create the dense relationship map from our DRU module. Additionally, we generated location relationship networks and semantic relationship networks based on the bounding box information and class information of the detected objects. For instance, in the scenario depicted in Fig. 7 (b), let's consider the relationship of the object labeled as 6. Semantically, this object is identified as a "car." Therefore, from the semantic relationships map, we can observe that the object 6 is connected with objects labeled as 3 and 4, both assigned a value of 1, as they also represent "cars". However, the semantic relationships map alone cannot distinguish the relationship difference with other objects. In our DRU relationships map, the object labeled as 6 exhibits strong connections with the ego-vehicle labeled as 0 and the people labeled as 10, and it also shows weak connections with other objects. It is observable that the relationship network provided by our DRU module is more comprehensive and contains richer information. However, as mentioned earlier, manually defined relationships are convenient for human understanding but may not necessarily


**FIGURE 7.** Visual ablation results for the DRU module. The first column display the results derived from the PE module after retaining the top 10 participants in the driving image. The object with the red rectangle is the most important object. The three columns on the right display the relationship networks acquired through our DRU module, the location relationship network based exclusively on bounding box information, and the semantic relationship network computed solely from object category information. In the three relationship heatmaps on the right, "0" corresponds to the ego-vehicle, while "1-10" represent the 10 objects detected from the scene. The complete heatmap illustrates the bidirectional relationship network among these 11 objects.

be suitable for machine learning. Our method, which allows the model to autonomously learn relationships and capture more nuanced relationship networks.

### 3) PERFORMANCE WITH DIFFERENT DRU LAYERS

Table 4 showcases the influence of varying the number of DRU layers in our DRUformer on the IOL task. We assessed the model's performance with 1 to 6 layers of DRU and monitored changes in ACC and mIoU. Notably, the optimal performance is attained with three layers. However, as the number of layers increases, the detection performance starts to deteriorate.

## V. DISCUSSION

Although our DRUformer model achieved excellent results in the IOL task, we did not conduct the following experiments due to limitations in computing hardware. In our experiments, the driving scene section only selected images from the last key frame of the driving video as the scene. However, in this scenario with a single image, we did not consider temporal information, focusing solely on spatial information. Even for a single image, our model already occupies 45GB of memory, while the A6000 graphics card has only 48GB of memory. However, the computational cost of this approach is slightly high for practical deployment scenarios. In the future, we plan to consider replacing the scene section with the entire video scene to enhance the temporal information of the PE module. Simultaneously, we intend to explore the addition of Optical Flow corresponding to the scene to enhance scene information. In addition, we consider also using more efficient Transformer implementations, such as EfficientFormerV2 [46] or similar alternatives.

The code of our paper and more detection results is available here: https://github.com/oniu-uin0/DRUformer.

## VI. CONCLUSION

To the best of our knowledge, DRUfomer model stands out as the first multi-modal model for important object detection in driving scenes based on the transformer architecture. This innovative approach considers both the driving scene and the driver's intention. Additionally, we introduce a driving scene relationship understanding module, specifically designed for machine comprehension, to enhance important object detection. This module eliminates the need for manually defining interactive relationships between objects. We conducted a comprehensive evaluation by comparing our method with other SOTA models on the largest important object detection dataset, DRAMA. The results of this comparison affirm the effectiveness of our model. Furthermore, we designed numerous ablation experiments to confirm the efficacy of our proposed intention extractor module and relationship understanding module.

While our research has successfully extracted driving intention commands from the driver, it currently encounters limitations in human-AV interaction. Effectively communicating with passengers or drivers is a critical area that needs improvement. In the next phase, our goal is to broaden our focus beyond the driver's intention commands to include more complex interactive commands. In scenarios where the driver lacks a specific driving intention, we acknowledge the

importance of driving intention and plan to adopt the path planning route as a representative driving intention in our research.

## REFERENCES

[1] A. Eskandarian, C. Wu, and C. Sun, "Research advances and challenges of autonomous and connected ground vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 683–711, Feb. 2021.

[2] California Department of Motor Vehicles. (2023). *Autonomous Vehicle Collision Reports, 2023*. Accessed: Sep. 23, 2023.

[3] *Road Safety*. Accessed: Sep. 23, 2023.

[4] Y. Niu, M. Ding, Y. Zhang, K. Ohtani, and K. Takeda, "Auditory and visual warning information generation of the risk object in driving scenes based on weakly supervised learning," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 1572–1577.

[5] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[6] S. Malla, C. Choi, I. Dwivedi, J. Hee Choi, and J. Li, "DRAMA: Joint risk localization and captioning in driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1043–1052.

[7] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Andrea Palazzi, France, Jun. 2016, pp. 54–60.

[8] A. Tawari, P. Mallela, and S. Martin, "Learning to attend to salient targets in driving videos using fully convolutional RNN," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3225–3232.

[9] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Proc. 14th Asian Conf. Comput. Vis.*, Perth, WA, Australia. Cham, Switzerland: Springer, 2019, pp. 658–674.

[10] P. V. Amadori, T. Fischer, and Y. Demiris, "HammerDrive: A task-aware driving visual attention model," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5573–5585, Jun. 2022.

[11] E. Ohn-Bar and M. M. Trivedi, "What makes an on-road object important?" in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3392–3397.

[12] M. M. Karim, R. Qin, and Z. Yin, "An attention-guided multistream feature fusion network for localization of risky objects in driving videos," 2022, *arXiv:2209.07922*.

[13] Z. Zhang, A. Tawari, S. Martin, and D. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8920–8927.

[14] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "HOTR: End-to-end human-object interaction detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 74–83.

[15] Y. Tian, A. Carballo, R. Li, and K. Takeda, "Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles," 2020, *arXiv:2011.13588*.

[16] S.-Y. Yu, A. V. Malawade, D. Muthirayan, P. P. Khargonekar, and M. A. A. Faruque, "Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7941–7951, Jul. 2022.

[17] H. Peng, F. Liu, Y. Li, B. Huang, J. Shao, N. Sang, and C. Gao, "Parallel reasoning network for human-object interaction detection," 2023, *arXiv:2301.03510*.

[18] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20072–20080.

[19] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, and J. Sun, "End-to-end human object interaction detection with HOI transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11820–11829.

[20] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 17853–17862.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[24] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[25] W. Wang, H. Bao, L. Dong, and F. Wei, "VLMo: Unified vision-language pre-training with mixture-of-modality-experts," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 32897–32912.

[26] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18113–18123.

[27] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16795–16804.

[28] Y. Niu, M. Ding, Y. Zhang, M. Ge, H. Yang, and K. Takeda, "Open-world driving scene segmentation via multi-stage and multi-modality fusion of vision-language embedding," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–6.

[29] Y. Niu, M. Ding, M. Ge, R. Karlsson, Y. Zhang, and K. Takeda, "R-cut: Enhancing explainability in vision transformers with relationship weighted out and cut," 2023, *arXiv:2307.09050*.

[30] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[31] T. Umeda, C. Miyajima, E. Takeuchi, and K. Takeda, "Modeling and evaluation of the gaze behavior of individual drivers," in *Proc. 8th Biennial Workshop Digit. Signal Process. In-Vehicle Mobile Syst.*, 2018, pp. 1–4.

[32] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8853–8859.

[33] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? Towards driver-centric risk assessment: Risk object identification via causal inference," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10711–10718.

[34] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10583–10591.

[35] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[36] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19526–19535.

[37] F. Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong, and S. Gould, "Exploring predicate visual context in detecting of human–object interactions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10411–10421.

[38] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, "GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20091–20100.

[39] Q. Wang, X. Guo, and H. Wang, "1st place solution for PSG competition with ECCV'22 SenseHuman workshop," 2023, *arXiv:2302.02651*.

[40] A. S. Kumar and J. J. Nair, "Scene graph generation using depth, spatial, and visual cues in 2D images," *IEEE Access*, vol. 10, pp. 1968–1978, 2022.

[41] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, "Mining the benefits of two-stage and one-stage hoi detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17209–17220.

[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[43] S. H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and A loss for bounding box regression," 2019, *arXiv:1902.09630*.
[44] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.
[45] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6748–6758.
[46] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, and J. Ren, "Rethinking vision transformers for MobileNet size and speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023.

**YINGJIE NIU** (Student Member, IEEE) received the B.S. degree in mechatronic engineering from China Three Gorges University, Yichang, China, in 2018, and the M.S. degree in mechatronic engineering from Southwest Jiaotong University, Chengdu, China, in 2021. He is currently pursuing the Ph.D. degree in intelligent systems with the Graduate School of Informatics, Nagoya University, Nagoya, Japan. His research interests include scene understanding, weakly supervised learning, and zero-shot learning.

**MING DING** (Member, IEEE) received the M.S. and Ph.D. degrees in engineering from Nara Institute of Science and Technology, Japan, in 2007 and 2010, respectively. In April 2010, he joined the Department of Mechanical Engineering, Tokyo University of Science, as a Postdoctoral Researcher. From October 2011 to February 2014, he was a Researcher with the RIKEN-TRI Collaboration Center for Human-Interactive Robot Research, RIKEN. Since March 2014, he has been a Designated Assistant Professor with the Graduate School of Engineering, Nagoya University, Japan. Since May 2015, he has been an Assistant Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. Since November 2019, he has been with the Institutes of Innovation for Future Society, Nagoya University, as a Designated Associate Professor. His current research interests include robot control, human modeling, and human–machine interface. He is a member of JSR.

**KEISUKE FUJII** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Kyoto University, in 2009, 2011, and 2014, respectively. After his work as a Postdoctoral Fellow and a Research Scientist with Nagoya University and the RIKEN Center for Advanced Intelligence Project in Japan, he joined Nagoya University, where he is currently an Associate Professor with the Graduate School of Informatics. His research interests include interdisciplinary studies among machine learning, behavioral sciences, and sports sciences.

**KENTO OHTANI** received the B.E., M.E., and Ph.D. degrees from Nagoya University, in 2013, 2015, and 2018, respectively. He is currently a Designated Assistant Professor with Nagoya University. His research interests include human behavior-related signal processing, including spatial audio and autonomous driving.

**ALEXANDER CARBALLO** (Member, IEEE) received the Dr.Eng. degree from the Intelligent Robot Laboratory, University of Tsukuba, Japan. From 1996 to 2006, he was a Lecturer with the School of Computer Engineering, Costa Rica Institute of Technology. From 2011 to 2017, he worked in LiDAR research and development with Hokuyo Automatic Company Ltd. In 2017, he joined Nagoya University as a Designated Associate Professor affiliated with the Institutes of Innovation for Future Society. In 2022, he was appointed as a permanent Associate Professor with the Graduate School of Engineering, Gifu University, Japan. His research interests include LiDAR sensors, robotic perception, and autonomous driving. He is a Professional Member of the IEEE Intelligent Transportation Systems Society (ITSS), the IEEE Robotics and Automation Society (RAS), the Robotics Society of Japan (RSJ), Asia–Pacific Signal and Information Processing Association (APSIPA), the Society of Automotive Engineers of Japan (JSAE), and Japan Society of Photogrammetry and Remote Sensing (JSPRS).

**KAZUYA TAKEDA** (Senior Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from Nagoya University, in 1983, 1985, and 1993, respectively. He has held positions at the Advanced Telecommunication Research Laboratories (ATR) and the KDD Research and Development Laboratory, in addition to being a Visiting Scientist with MIT, before rejoining Nagoya University, in 1995. He is currently the Vice President of Nagoya University and a Professor with the Institute of Innovation for Future Society and the Graduate School of Informatics, Nagoya University. He is also the Director of Tier IV Inc. He has co-founded Tier IV, a university startup aimed at democratizing autonomous driving technologies through the development of the open-source software platform, Autoware. His research interests include signal processing and machine learning of behavior signals and their applications. With over 150 journal articles, nine coauthored/co-edited books, and 15 patents to his name, he is a prolific contributor to his field. From 2013 to 2022, he was a Board of Governors Member of the IEEE ITS Society and the Asia–Pacific Signal and Information Processing Association (APSIPA). He is a Governors Member of the IEEE ITS Society and APSIPA. He is a fellow of IEICE. His achievements include the 2020 IEEE ITS Society Outstanding Research Award and six best paper awards from IEEE international conferences and workshops, in addition to various domestic awards. He chaired several scientific meetings, including FAST-Zero, in 2017, and Universal Village, in 2016. He served as the Program Chair for IEEE ICVES 2009 and IEEE ITSC 2017. Furthermore, he was the General Chair of the IEEE Intelligent Vehicle Symposium (IV2021).