On-Road Object Importance Estimation: A New Dataset and A Model with Multi-Fold Top-Down Guidance

Zhixiong Nan[1], Yilong Chen[1], Tianfei Zhou [*2], and Tao Xiang[1]

[1]College of Computer Science, Chongqing University, Chongqing, China.
[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.
nanzx@cqu.edu.cn, chenyilong@stu.cqu.edu.cn, tfzhou@bit.edu.cn,
txiang@cqu.edu.cn [†]

Abstract

This paper addresses the problem of on-road object importance estimation, which utilizes video sequences captured from the driver's perspective as the input. Although this problem is significant for safer and smarter driving systems, the exploration of this problem remains limited. On one hand, publicly-available large-scale datasets are scarce in the community. To address this dilemma, this paper contributes a new large-scale dataset named Traffic Object Importance (TOI). On the other hand, existing methods often only consider either bottom-up feature or single-fold guidance, leading to limitations in handling highly dynamic and diverse traffic scenarios. Different from existing methods, this paper proposes a model that integrates multi-fold top-down guidance with the bottom-up feature. Specifically, three kinds of top-down guidance factors (*i.e.*, driver intention, semantic context, and traffic rule) are integrated into our model. These factors are important for object importance estimation, but none of the existing methods simultaneously consider them. To our knowledge, this paper proposes the first on-road object importance estimation model that fuses multi-fold top-down guidance factors with bottom-up feature. Extensive experiments demonstrate that our model outperforms state-of-the-art methods by large margins, achieving **23.1**% Average Precision (AP) improvement compared with the recently proposed model (*i.e.*, Goal).

1 Introduction

According to the World Health Statistics of WHO ([34]), road traffic injuries account for a significant 29% of all injury deaths, with nearly 1.3 million people losing their lives in traffic accidents annually. Accurately estimating the importance of on-road objects can pave the way for safer (*e.g.*, automatic emergency braking ([45]; [39])) and smarter driving systems ([32]; [25]; [38]; [49]; [4]; [36]; [11]), potentially preventing numerous accidents.

Although on-road object importance estimation presents significant research value, it has not been widely explored. One main reason is the scarcity of publicly-available large-scale datasets in the community. Specific for the on-road object importance estimation task, the only one publicly-available dataset is Ohn-Bar *et al.* ([33]), which contains 3,187 frames, 8 scenes, and 16,076 object importance annotations. Such small-scale dataset supports to train small and less complex models.

---

[*]Corresponding author.

However, traffic scenes are dynamic and diverse, asking for complex models to handle various traffic situations. Some researchers have recognized this dilemma and propose some datasets (8; 21; 46). Unfortunately, these dataset are not publicly-available, thereby the dilemma has not been fundamentally addressed. To fundamentally address this dilemma and push forward the advancement of on-road object importance study, this paper will release a large-scale dataset (named as **TOI**, Traffic Object Importance) containing 9,858 frames, 28 scenes, and 44,120 object importance annotations. Compared to Ohn-Bar (33), **TOI** achieves a 3.1-fold increase in frames count, a 3.5-fold increase in scene count, and a 2.7-fold increase in object count.

From the perspective of methodology, some methods have been proposed (21; 8; 50; 33). However, these methods are relatively simple, exhibiting the low performance when confronting to challenging traffic scenarios. This motivates us to think about a question: why can not existing methods perform well? The potential reason is that existing on-road importance estimation methods underestimate the complexity of traffic scenarios, individual bottom-up mechanism (50) (assuming important objects are the objects with salient color, texture, size, *etc.*) or simple top-down guidance mechanism (8; 21) (fusing the bottom-up information with a certain type of top-down information such as semantics, ego-car trajectory, driving task (27), *etc.*) can hardly address dynamic and diverse scenarios.
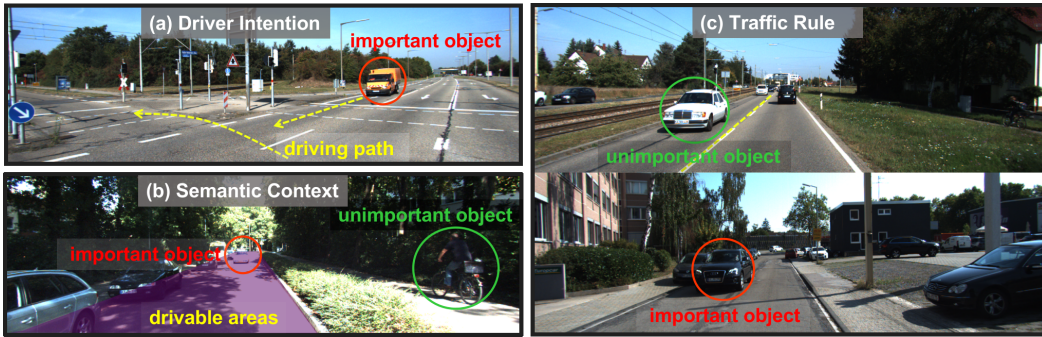

Figure 1: The crucial factors considered by human drivers when estimating on-road object importance.

Therefore, a smarter model is needed. Inspired by the fact that a human driver can accurately estimate object importance in challenging situations, this paper attempts to design a model by analysing the human reasoning mechanism during estimating object importance. To this end, the primary question is "*what essentially crucial factors are considered when a human driver is estimating the importance of objects?*". **Firstly**, *the attributes (e.g., size, color, and texture) of the object* is considered. For example, when a truck with the big size and a car with the small size simultaneously appear in front of the ego-car, the truck is more important since it imposes bigger impact on the driving. **Secondly**, *the driver intention* is considered. The objects that will riskly collide with the ego-car intention driving path or the objects locating on the ego-car intention driving path present high importance, as shown in Fig. 1a. **Thirdly**, *overall semantic context of the whole traffic scene* is considered. A human usually pay more attention on the objects in drivable areas rather than the objects in undrivable areas. As shown in Fig. 1b, the person riding a bicycle in undrivable areas is unimportant. **Fourthly**, *traffic rule* is considered. Most of traffic participants obey the traffic rule, thus the traffic rule is an critical factor for a human to estimate object importance. For example, as shown in Fig. 1c, when there exists a lane marking between the oncoming car and the ego-car, the human driver may consider the oncoming car as unimportant. In contrast, if there is no lane marking between them, the importance of the oncoming car significantly increases. The traffic rule is crucial for object importance estimation. However, none of existing methods utilizes traffic rule to estimate on-road object importance.

Based on the above observations, we propose a model with multi-fold top-down guidance including *driver intention*, *semantic context*, and *traffic rule*. As far as we know, it is the first on-road object importance estimation model that fuses multi-fold top-down guidance factors with the bottom-up feature. The proposed model consists of two kinds of pathways (*i.e.*, bottom-up pathway and top-down pathway). Bottom-up pathway and top-down pathway are fused to estimate object importance. Specifically, in the top-down pathway, the top-down guidance factors of *driver intention* and *semantic context* are involved in the proposed Driver Intention and Semantics Guidance (**DISG**) module, and *traffic rule* is modeled in the proposed Traffic Rule Guidance (**TRG**) module. In the bottom-up pathway, Object Feature Extraction (**OFE**) module is proposed to extract object features in both spatial and temporal dimensions.

2

A series of comparison and ablation studies are conducted on a public dataset (33) and our **TOI** dataset. The comparison experiment results show that our model has a solid advantage over the baselines. The ablation study results validate the effectiveness of our proposed interactive bottom-up & top-down fusion framework and multi-fold top-down guidance modules (*i.e.*, **DISG** and **TRG**).

The contributions of this paper are as follows. **1)** This paper contributes a new large-scale dataset, which will be publicly released. This dataset is almost three times larger than the current publicly available public dataset (33). **2)** This paper contributes an object importance estimation model. As far as we know, it is the first on-road object importance estimation model that integrates multi-fold top-down guidance factors with the bottom-up feature. **3)** The traffic rule is crucial for object importance estimation. However, none of existing methods utilizes traffic rule to estimate on-road object importance. This paper considers the effect of traffic rule on object importance and successfully models this abstract concept by proposing an adaptive object-lane interaction mechanism.

2 Related Works

On-Road Object Importance Estimation Related Datasets. Currently, the primary dilemma of research on the on-road object importance estimation is the lack of sufficient data. Among existing datasets relevant to autonomous driving perception tasks, only a few meet the data requirement of providing images from the driver's first-person perspective while also including object importance level labels. Ohn-Bar *et al.* (33) are the first to define the problem of on-road object importance estimation, and they propose a small-scale publicly available dataset, which contains 8 scenes. Gao *et al.* (8) and Li *et al.* (21) have significantly increased the number of scenes. However, their datasets are not publicly available, thus the contributions to the community is limited. Datasets (19; 41; 48) are with detailed annotations such as ego's reaction and road topology. They have great potential to advance the development of on-road object importance estimation. However, currently, they lack object importance level labels and cannot be directly applied to this task.

On-Road Object Importance Estimation Related Methods. Currently, on-road object importance estimation methods can be divided into two categories: 1) methods solely utilizing bottom-up feature; 2) methods utilizing single-fold top-down guidance.

The methods solely utilizing bottom-up feature focus on the visual attributes of the objects. The bottom-up processing method is initially introduced in (44), and Itti *et al.* (17) propose one of the first bottom-up mechanism based models. Following this, many researchers are inspired by this concept (43; 15; 18; 35). Zhang *et al.* (50) introduce a model, which solely relies on RGB clips for object importance estimation. This model employs graph convolutions to characterize the interactions among on-road objects. Nitta *et al.* (30) develop a model that extracts temporal features from optical flow images to infer the states of moving objects. The optical flow images are also used in Malla *et al.* (26) to assess the states of moving objects. The bottom-up methods can also be found in the works (33; 52; 47; 23; 14; 24). Although the bottom-up feature is crucial for importance estimation, the methods solely rely on bottom-up feature can not function well in the complex scenarios.

The importance of an object is influenced by many factors such as driver intention, which cannot be fully utilized through bottom-up methods. However, current methods are relatively simple and relies on single-fold guidance. Niu *et al.* (31) utilize a Transformer with shared weights to identify high-risk objects and generated semantic warning sentences. Li *et al.* (21) investigate the impact of driver intention, employing the action and trajectory data of the ego-car as additional supervisory signals in auxiliary tasks to enhance model performance. Gao *et al.* (8) utilize the driver's goal to estimate object importance. A cause-effect problem was formulated in (20), which introduced a model to estimate the risky object by considering its potential impact on the driver's behavior. Tang *et al.* (42) provide a more comprehensive understanding of how driver intentions under different tasks affect the driver attention. Single-fold top-down guidance can also be found in the works of attention prediction task (7; 16; 22; 6; 1; 29; 5; 28). However, none of these methods utilizes multi-fold top-down guidance factors to estimate the on-road object importance.

3 A New Dataset: TOI

We thoroughly review existing datasets for on-road object importance estimation as well as the datasets for the related tasks, and provide a summary in Tab. 1. The datasets for the related tasks (*e.g.*, risk assessment (37; 48; 19), accident anticipation (41), and situation awareness (40)) do not include object importance labels, making them unsuitable for object importance estimation task. Most datasets (*e.g.*, (21; 8)) for object importance estimation are not publicly available. The only

Table 1: Comparison between the TOI and State-of-the-art Datasets. 'Impo.' represents the object importance annotation.

Dataset	Task	Public	Impo.	Extra-Information GPS/IMU	Lidar	3D-Labels	Objects	Frames	Scenes	FPS	Year
HDD (37)	risk assessment	✓	✗	✓	✓	✗	-	-	-	30	2018
1361-honda (48)	risk assessment	✓	✗	✗	✗	✗	-	-	1,361	-	2020
RiskBench (19)	risk assessment	✓	✗	✗	✗	✗	-	-	6,916	-	2024
NIDB (41)	accident anticipation	✓	✗	✗	✗	✗	-	499,500	4,995	-	2018
A-SASS (46)	situation awareness	✗	✓	✗	✗	✗	-	-	10	30	2022
ROAD (40)	situation awareness	✓	✗	✓	✓	✓	560,000	122,000	22	12	2023
Ohn-Bar (33)	on-road object importance estimation	✓	✓	✓	✓	✓	16,076	3,187	8	10	2017
Goal (8)	on-road object importance estimation	✗	✓	✓	✗	✗	-	244,980	743	30	2019
Li (21)	on-road object importance estimation	✗	✓	✓	✓	✓	-	-	-	2	2022
TOI	on-road object importance estimation	✓	✓	✓	✓	✓	44,120	9,858	28	10	2024

publicly available dataset is Ohn-Bar (33), but it is a small scale dataset. In response to the scarcity of publicly-available large-scale datasets for on-road object importance estimation, we contribute a large-scale dataset named **TOI** (Traffic Object Importance). **TOI** is built by re-annotating the authoritative KITTI (9) dataset. While there are many datasets (*e.g.*, nuScenes (2; 40; 37)) that be used for object importance annotations, we select KITTI dataset for the following reason: KITTI is the worldwide benchmark in the field of autonomous driving. In addition, KITTI is collected in diverse real traffic scenes including rural areas and on highways with rich date formats, making the dateset friendly for various tasks.

Annotation Procedure. The criterion of object importance might be ambiguous as different drivers usually hold different opinions on the importance judgment. Currently, object-level importance labels are annotated without checking mechanism. Although multiple annotators perform the annotations, the annotations finished by the certain annotator are not checked by others, leading to the unreliable and ambiguous annotations. To generate reliable annotations, we adopt two mechanisms: the *double-checking annotation mechanism* and the *triple-discussing annotation mechanism*.

The *double-checking annotation mechanism* operates as follows. Initially, the first annotator (an experienced driver) labels the object importance at every 10 frames. To guarantee the reliability of annotations, the first annotator only selects one object as the important object at each time of observing the whole 10 frames. The annotation for these 10 frames is finished until all important objects are annotated, then the annotator moves to the next set of 10 frames for annotation. Subsequently, the annotation results are checked by the second annotator (who is also an experienced driver). When the second annotator finds a disputed annotation, the first and second annotators discuss together to reach an agreement. If they are unable to reach an agreement, the *triple-discussion annotation mechanism* is activated. In this case, the third annotator is invited to discuss the final annotations.

Statistics and Comparison. Totally, 9,858 image frames are annotated, generating 44,120 objects with the importance or unimportance annotations, among which 5,052 objects are annotated as important. The annotated data are randomly split into training/testing datasets with a ratio of 8,121 : 1,737. The comparison between **TOI** and existing on-road object importance estimation datasets and similar task datasets are presented in Tab. 1. Compared to the publicly available Ohn-Bar (33) dataset, **TOI** represents the significant advantages in following three aspects. **Frame quantity**: **TOI** exhibits a substantial increase in the number of frames, with 9,858 frames compared to 3,187 frames in the Ohn-Bar dataset. **Object quantity**: the number of annotated objects is 44,120 compared to 16,076 in the Ohn-Bar dataset. **Scene diversity**: while Ohn-Bar contains only 8 scenes, **TOI** covers 28 scenes. Compared to Goal (8) dataset, **TOI** has rich annotations such as Lidar and 3D tracklet labels. The abundance of multimodal annotations enables **TOI** to support the research on on-road object importance estimation using multimodal learning methods in the future. Though Goal (8) presents the advantage in terms of frame number and scene diversity, it is not publicly available. Compared to Li dataset (21), **TOI** offers the frame rate of 10 FPS. This high temporal resolution is critical for capturing the dynamic changes of on-road object importance.

Annotation Challenges. Compared to datasets for other tasks, **TOI** may not be considered large-scale, it is relatively large-scale compared to existing publicly available datasets for on-road object importance estimation. However, annotating object importance at this scale is challenging. Each annotation requires multiple annotators and undergoes rigorous checking to achieve satisfactory results. In addition, to generate reliable annotations, only one object is annotated at each time observing the whole video sequence, a complete re-observation of the whole video sequence is required for the annotation of the next object. Moreover, object importance is affected by multiple factors, which imposes difficulties on the annotation.
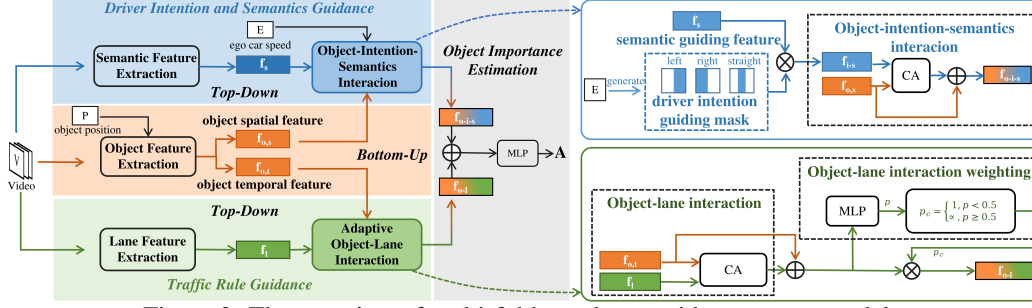
4 Approach

4.1 Overview

Figure 2: The overview of multi-fold top-down guidance aware model.

Consider a traffic scenario with N on-road objects in T time steps, the goal of this work is to estimate on-road object importance (A) at the final time step (*i.e.*, $t = T$) using the video sequence (V) captured from the driving perspective over T time step and ego-car velocity information (E) at the first time step (*i.e.*, $t = 1$), which is formulated as:

$$A = \mathcal{N}(V, E), \tag{1}$$

where $\mathcal{N}$ represents an on-road object importance estimation network, $V = \{V_t\}_{t=1}^T$, and $A = \{A_i\}_{i=1}^N$.

In order to effectively fuse multi-fold top-down guidance factors (*i.e.*, *semantic context*, *driver intention*, and *traffic rule*) with bottom-up object visual feature, we propose a multi-fold top-down guidance aware model, the overview of which is illustrated in Fig. 2. Our model is composed of four key modules: **Object Feature Extraction** (**OFE**) module detailed in § 4.2, **Driver Intention and Semantics Guidance** (**DISG**) module described in § 4.3, **Traffic Rule Guidance** (**TRG**) module explained in § 4.4, and **Object Importance Estimation** module introduced in § 4.5.

Firstly, **OFE** extracts object spatial feature $f_{o,s}$ and object temporal feature $f_{o,t}$ from V. Then, **DISG** takes E, V, and $f_{o,s}$ as inputs, and outputs object-intention-semantics interaction feature $f_{o\text{-}i\text{-}s}$. Meanwhile, in **TRG**, the lane feature f_l and $f_{o,t}$ are processed by **adaptive object-lane interaction** mechanism to produce the object-lane interaction feature $f_{o\text{-}l}$. Finally, $f_{o\text{-}i\text{-}s}$ and $f_{o\text{-}l}$ are used to estimate object importance A.

4.2 Object Feature Extraction

The goal of **Object Feature Extraction** (**OFE**) module is to extract object features in both temporal and spatial dimensions. The input of **OFE** is a RGB video sequence $V \in \mathbb{R}^{T \times 3 \times W \times H}$, and the outputs are object temporal feature $f_{o,t}$ and object spatial feature $f_{o,s}$.

To begin with, **OFE** takes V and M (M denote optical flow images derived from V) as inputs to extract the object visual feature $f_v \in \mathbb{R}^{N \times T \times C \times W' \times H'}$ (reflecting the appearance of the object) and the object motion feature $f_m \in \mathbb{R}^{N \times T \times C \times W' \times H'}$ (reflecting the movement of the object). This procedure is formulated as:

$$f_v = \text{Roi}(\mathcal{N}_V(V)), \quad f_m = \text{Roi}(\mathcal{N}_M(M)), \tag{2}$$

where $\mathcal{N}_V$ and $\mathcal{N}_M$ represent the two ResNet18 (12), Roi denotes the ROI pooling (10), C represents the number of channels, W' and H' denote the width and height obtained through ROI pooling.

Subsequently, object spatial feature $f_{o,s} \in \mathbb{R}^{N \times 2C \times W' \times H'}$ is obtained based on f_v and f_m. The goal of $f_{o,s}$ is to focus on the spatial information of objects. Therefore, an average pooling is applied on the time dimension (*i.e.*, the dimension of T) of f_v and f_m, and a self-attention mechanism is utilized to emphasize the spatial information (*i.e.*, the dimensions of W' and H'), which is denoted as follows:

$$f_{o,s} = \mathcal{N}_{mhsa}(\text{Concat}(\text{Avg}(f_v), \text{Avg}(f_m))), \tag{3}$$

where Avg denotes average pooling, Concat is concatenation. $\mathcal{N}_{mhsa}$ represents the multi-head self-attention mechanism, and it has the same meaning in the following parts.

Meanwhile, object temporal feature $\boldsymbol{f}_{o,t} \in \mathbb{R}^{N \times C'}$ is also extracted based on $\boldsymbol{f}_v$ and $\boldsymbol{f}_m$. The goal of $\boldsymbol{f}_{o,t}$ is to focus on the temporal feature of objects. Therefore, the dimensions of $\boldsymbol{f}_v$ and $\boldsymbol{f}_m$ are firstly reshaped from $N \times T \times C \times W' \times H'$ to $N \times T \times (C \times W' \times H')$, and then two LSTM networks are applied to extract the temporal feature. Finally, to subsequently fuse $\boldsymbol{f}_{o,t}$ with $\boldsymbol{f}_l \in \mathbb{R}^{N \times C'}$, it is necessary to transform the channel number of $\boldsymbol{f}_{o,t}$, hence a Linear layer is required. This procedure is formulated as:

$$\boldsymbol{f}_{o,t} = \text{Linear}(\mathcal{N}_{mhsa}(\text{Concat}(\mathcal{N}_{lstm}(\boldsymbol{f}_v), \mathcal{N}_{lstm}(\boldsymbol{f}_m)))), \tag{4}$$

where $\mathcal{N}_{lstm}$ is the LSTM network (13), which outputs features for T time steps, and $\boldsymbol{f}_{o,t}$ is computed by indexing the information at the T-th time step.

4.3 Driver Intention and Semantics Guidance

The driver intention and semantic context significantly affect the on-road object importance in a top-down manner, thus we propose the **Driver Intention and Semantics Guidance** (**DISG**) module. **DISG** consists of two components, namely **semantic feature extraction** and **object-intention-semantics interaction**. The former extracts the *semantic guiding feature* (*i.e.*, $\boldsymbol{f}_s$ in Eq. (5)) from driving scenarios, and the latter uses $\boldsymbol{f}_s$ and *intention guiding mask* (*i.e.*, $\boldsymbol{m}$ in Eq. (7)) to guide the refinement of $\boldsymbol{f}_{o,s}$. In the following part, we will sequentially introduce the calculation processes for *semantic guiding feature* and *intention guiding mask*.

Semantic guiding feature. The goal of $\boldsymbol{f}_s \in \mathbb{R}^{1 \times 2C \times W' \times H'}$ is to guide the model to be aware of the semantic context of on-road objects, which is extracted by:

$$\boldsymbol{f}_s = \mathcal{N}_G(\boldsymbol{G}), \tag{5}$$

where $\mathcal{N}_G$ denotes a ResNet18 backbone network and $\boldsymbol{G}$ represents semantic segmentation maps obtained from $\boldsymbol{V}_T$.

Intention guiding mask. The goal of $\boldsymbol{m} \in \mathbb{R}^{W' \times H'}$ is making the model be aware of the region that is consistent with the driver intention. However, it is difficult to realize since the driver intention is an abstract concept and the limited information regarding the driver intention is known. Additionally, it is not reasonable to assume the driver intention is known in advance. Therefore, we use three common intention behaviors in driving to reflect the driver intention (*i.e.*, *turning left*, *going straight*, and *turning right*). Intention behaviors are formulated as the corresponding intention guiding masks:

$$\boldsymbol{m}_l = \begin{bmatrix} a & \dots & a & b & \dots & b \\ a & \dots & a & b & \dots & b \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a & \dots & a & b & \dots & b \end{bmatrix}_{(W' \times H')}, \boldsymbol{m}_s = \begin{bmatrix} a & \dots & a & b & \dots & b & a & \dots & a \\ a & \dots & a & b & \dots & b & a & \dots & a \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a & \dots & a & b & \dots & b & a & \dots & a \end{bmatrix}_{(W' \times H')}, \boldsymbol{m}_r = \begin{bmatrix} b & \dots & b & a & \dots & a \\ b & \dots & b & a & \dots & a \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b & \dots & b & a & \dots & a \end{bmatrix}_{(W' \times H')} \tag{6}$$

where $\boldsymbol{m}_l$, $\boldsymbol{m}_s$, and $\boldsymbol{m}_r$ are manually engineered masks to emphasize the information in the right, center, and left regions of the images in the video sequence, respectively. Their sizes are aligned with the size of $\boldsymbol{f}_s$. a and b represent pre-determined low and high values, respectively. We note that $\boldsymbol{m}_l$ representing turning left behavior is allocated with higher value at the right part, which is inspired by the finding of Tang *et al.* (42) demonstrating that when a car is turning left, the driver pays more attention to the right side. Before proposing this predefined intention guiding mask, we design a learnable mask with random initialization to make the model automatically learn the mask to reflect intention behaviors. However, this strategy does not work. The potential reason is that the intention is abstract to learn.

To automatically select the $\boldsymbol{m}$ of corresponding driving behavior, we design the following logic based on the angular velocity E of ego-car:

$$\boldsymbol{m} = \begin{cases} \boldsymbol{m}_l, & \text{if } E > \beta \\ \boldsymbol{m}_r, & \text{if } E < -\beta \\ \boldsymbol{m}_s, & \text{otherwise} \end{cases}, \tag{7}$$

where β represents the driver turning threshold.

After obtaining $\boldsymbol{f}_s$ and $\boldsymbol{m}$, **DISG** module fuses them and uses the fused feature to guide the refinement of $\boldsymbol{f}_{o,s}$. This procedure is implemented by the **object-intention-semantics interaction** component. The first task of **object-intention-semantics interaction** is to fuse $\boldsymbol{f}_s$ and $\boldsymbol{m}$, and generate the intention-semantics interaction feature $\boldsymbol{f}_{i\text{-}s} \in \mathbb{R}^{1 \times 2C \times W \times H}$, which is denoted as follow:

$$\boldsymbol{f}_{i\text{-}s} = \boldsymbol{f}_s \times \boldsymbol{m}. \tag{8}$$

where the operator $\times$ makes the model pay more attention to the semantic context in the driver intention regions.

The second task of **object-intention-semantics interaction** is to refine $\boldsymbol{f}_{o,s}$ by interacting with $\boldsymbol{f}_{i\text{-}s}$, which is formulated as follow:

$$\boldsymbol{f}_{o\text{-}i\text{-}s} = \mathcal{N}_{mhca}(\boldsymbol{f}_{o,s}, \boldsymbol{f}_{i\text{-}s}) + \boldsymbol{f}_{o,s}, \tag{9}$$

where $\mathcal{N}_{mhca}$ denotes the multi-head cross-attention mechanism, $\boldsymbol{f}_{o,s}$ serves as the *query* while $\boldsymbol{f}_{i\text{-}s}$ serves as the *key* and *value*, and $\boldsymbol{f}_{o\text{-}i\text{-}s} \in \mathbb{R}^{N \times 2C \times W' \times H'}$.

4.4 Traffic Rule Guidance

On-road object importance is also closely related with traffic rule, but it is often overlooked in previous works. To effectively leverage the traffic rule, we propose the **Traffic Rule Guidance** (**TRG**) module, which consists of two components: **lane feature extraction** and **adaptive object-lane interaction**.

Lane feature extraction is to make the preparation for **adaptive object-lane interaction**. In detail, a linear transformation and an activation are applied on lane information $\boldsymbol{L}$:

$$\boldsymbol{f}_l = \text{Relu}(\text{Linear}(\boldsymbol{L})), \tag{10}$$

where $\boldsymbol{L}$ are the coordinates of lane marking points, which are derived from $\boldsymbol{V}_T$ via a lane marking detector, Relu is the rectified linear unit activation, and $\boldsymbol{f}_l \in \mathbb{R}^{N \times C'}$.

Adaptive object-lane interaction is the core of **TRG**, and it contains two steps: *object-lane interaction* and *object-lane interaction weighting*. In the first step, lane feature $\boldsymbol{f}_l$ and $\boldsymbol{f}_{o,t}$ are fused through a multi-head cross-attention mechanism and a residual mechanism, which can be denoted as:

$$\boldsymbol{f}_{o\text{-}l}^m = \mathcal{N}_{mhca}(\boldsymbol{f}_l, \boldsymbol{f}_{o,t}) + \boldsymbol{f}_{o,t}, \tag{11}$$

where $\boldsymbol{f}_{o\text{-}l}^m \in \mathbb{R}^{N \times C'}$ denotes initial object-lane interaction feature, and $\boldsymbol{f}_l$ serves as the *query* while $\boldsymbol{f}_{o,t}$ serves as the *key* and *value*,

Factually, $\boldsymbol{f}_{o\text{-}l}^m$ has considered the traffic rule factor by modeling the relation between lane markings and on-road objects. However, the influence of lane markings on on-road object importance estimation might not be universally-effective in all scenarios, thus we propose the *object-lane interaction weighting* mechanism to realize **adaptive object-lane interaction**.

The goal of *object-lane interaction weighting* is to adaptively penalize the cases in which object-lane relation is weak (*e.g.*, static roadside cars weakly interacts with lane markings). To this end, a MLP network is applied on $\boldsymbol{f}_{o\text{-}l}^m$ to compute a score p, and this score is then used to compute the corresponding penalizing coefficient p_c, which is denoted as:

$$p = \text{Sigmoid}(\mathcal{N}_{mlp}(\boldsymbol{f}_{o\text{-}l}^m)), \tag{12}$$

$$p_c = \begin{cases} 1, & \text{if } p < 0.5 \\ \alpha, & \text{if } p \geq 0.5 \end{cases}, \tag{13}$$

where α is a very small value.

Based on p_c, the object-lane interaction feature $\boldsymbol{f}_{o\text{-}l} \in \mathbb{R}^{N \times C'}$ is obtained by weighting $\boldsymbol{f}_{o\text{-}l}^m$ in Eq. (11), which is denoted as:

$$\boldsymbol{f}_{o\text{-}l} = \boldsymbol{f}_{o\text{-}l}^m \times p_c. \tag{14}$$

We note that *object-lane interaction weighting* is the core of **adaptive object-lane interaction**, which is significant for object importance estimation (30.4% improvement on AP).

4.5 Object Importance Estimation

Taking $\boldsymbol{f}_{o\text{-}i\text{-}s}$ in Eq. (9) and $\boldsymbol{f}_{o\text{-}l}$ in Eq. (14) as the inputs, object importance $\boldsymbol{A} \in \mathbb{R}^N$ is estimated. This procedure is formulated as:

$$\boldsymbol{A} = \text{Softmax}(\mathcal{N}_{mlp}(\text{Linear}(\boldsymbol{f}_{o\text{-}i\text{-}s}) + \boldsymbol{f}_{o\text{-}l})), \tag{15}$$

where $\boldsymbol{A}$ signifies the importance for each object. The Linear layer transforms the dimensions of $\boldsymbol{f}_{o\text{-}i\text{-}s}$ from $N \times 2C \times W' \times H'$ to $N \times C'$ so that it can be added to $\boldsymbol{f}_{o\text{-}l}$.

5 Experiments

Metrics. For performance evaluation, two classical metrics are chosen: Average Precision (AP) and F1 Score (F1). AP is computed by calculating the area under the precision-recall curve at various thresholds, thus it is a compressive metric to indicate both the precision and recall of a model. F1 is computed by precision and recall at a fixed threshold, thus it indicates the balance between precision and recall. Both metrics follow the principle that higher values indicate better performance.

Loss Function. The loss function is defined as follow:

$$\mathcal{L}(\hat{A}, A) = \text{BCELoss}(\hat{A}, A) + \text{FocalLoss}(\hat{A}, A), \tag{16}$$

where A is the predicted object importance, $\hat{A}$ is the ground-truth.

5.1 Comparison Experiment

Baselines. Our model is compared with seven models. Ohn-Bar (33), Goal (8), Zhang (50), and Li (21) are representative works for on-road object importance estimation. In addition, considering that the salient object detection indicates important objects to the certain extent, three recently-proposed salient object detection models, namely MENet (23), A2S (52), and PGNet (47), are also selected as baselines.

Table 2: Quantitative comparison with baselines on TOI and Ohn-Bar (33) datasets. The 'Video' signifies the usage of RGB video sequence, 'Velocity' denotes the incorporation of vehicle velocity information, and '3D-Object' indicates the utilization of 3D object properties information.

Method	Inputs			TOI			Ohn-Bar			Speed (ms/clip)
	Video	Velocity	3D-Object	AP↑	F1↑	Acc↑	AP↑	F1↑	Acc↑	
Yolo				5	9	14	25	48	34	278
Ohn-Bar (33) PR'2017	✓	✓	✓	19	28	74	39	14	62	100
Goal (8) ICRA'2019	✓	✓		50	49	90	52	26	65	155
Zhang (50) ICRA'2020	✓			16	0	91	28	10	54	200
Li (21) ICRA'2022	✓	✓		23	0	91	41	0	64	202
MENet (23) CVPR'2023	✓			-	26	76	-	12	62	122
A2S (52) CVPR'2023	✓			-	9	78	-	18	61	120
PGNet (47) CVPR'2022	✓			-	30	84	-	27	55	130
Ours	✓	✓		**60**	**54**	**92**	**64**	**53**	**69**	115

Quantitative Comparison. Tab. 2 shows the comparison results on **TOI** and Ohn-Bar (33) datasets. We can observe that our model outperforms all seven baselines. On the Ohn-Bar (33) dataset, our model achieves **23.1%** and **96.3%** performance improvements on AP and F1 metrics compared with the second-best result, respectively. On the **TOI** dataset, compared with the second-best result on AP and F1 metrics, our model achieves **20.0%** (*i.e.*, (60-50)/50) and **10.2%** performance improvements, respectively. The results of MENet (23), A2S (52), and PGNet (47) are not reported on AP metric due to the lack of confidence scores in salient object detection outputs, making it impossible to calculate precision-recall curves for different thresholds. The F1 results for Zhang (50) and Li (21) are both 0 because they predict all objects as unimportant.

5.2 Ablation Studies

Top-down and Bottom-up Framework. To validate the effectiveness of our model that interactively integrates multi-fold top-down guidance mechanisms with bottom-up features, we conduct four experiments: **#1:** only the bottom-up module is enabled; **#2:** the bottom-up module is combined with **TRG** module; **#3:** the bottom-up module is combined with **DISG** module; **#4:** the bottom-up module is combined with both **TRG** module and **DISG** module. The experimental results are reported in Tab. 3. Compared with **#1**, **#2** achieves 55% and 40% performance improvements on AP and F1, respectively. Similarly, **#3** obtains 160% and 56% performance improvements on AP and F1, respectively. When both modules are enabled (**#4**), our model exhibits the best performance. These results validate both **TRG** and **DISG** modules are effective.

Table 3: Ablation study on top-down and bottom-up framework.

Method	BU	TRG	DISG	AP↑	F1↑
#1	✓			20	25
#2	✓	✓		31	35
#3	✓		✓	52	39
#4	✓	✓	✓	60	54

Driver Intention and Semantics Guidance (**DISG**). To verify the effectiveness of semantic context guidance and driver intention guidance, we conduct three experiments: **#1**: the model without semantic guiding feature and intention guiding mask; **#2**: the model only with semantic guiding feature; **#3**: the model only with intention guiding mask; **#4**: the model with both semantic guiding feature and intention guiding mask.

Table 4: Ablation study on DISG.

Method	Seman.	Intent.	AP↑	F1↑
#1			31	35
#2	✓		49	48
#3		✓	35	39
#4	✓	✓	**60**	**54**

The results are shown in Tab. 4. Compared to **#1**, **#2** yields 58.1% and 37.1% performance improvements on AP and F1 metrics, respectively. This enhancement is attributed to the usage of the semantic guiding feature, which enables the model to learn the semantic relation between objects and the whole scene. Meanwhile, **#3** achieves 15.4% and 11.4% performance improvements on AP and F1 metrics, respectively. The effectiveness of our intention guiding mask accounts for this advancement. **#4** exhibits the best performance, demonstrating the effectiveness of our proposed semantic guiding feature and intention guiding mask.

Traffic Rule Guidance (**TRG**). To analyze the effects of *object-lane interaction* and *object-lane interaction weighting*, we conduct three experiments: **#1**: the model without *object-lane interaction* and *object-lane interaction weighting*; **#2**: only the *object-lane interaction* is enabled; **#3**: both the *object-lane interaction* and the *object-lane interaction weighting* are enabled.

Table 5: Ablation study on TRG.

Method	Interac.	Weight.	AP↑	F1↑
#1			52	39
#2	✓		46	45
#3	✓	✓	**60**	**54**

The corresponding results are summarized in Tab. 5. Compared to **#1**, **#3** obtains 15.4% and 38.5% improvements on the AP and F1 metrics, respectively. These results demonstrate the significance of both *object-lane interaction* and the *object-lane interaction weighting* mechanisms. The reason is explainable. When lane information is not utilized, the implicit traffic rule conveyed by the lane is not used. The absence of the traffic rule results in a reduced ability of the model.

It comes as a surprise that the individual usage of *object-lane interaction* (**#2**) leads to the performance decreasing on the AP metric compared to **#1**. This is due to the fact that not all on-road objects are influenced by lanes. Without *object-lane interaction weighting*, individual *object-lane interaction* generates a uniform object-lane interaction feature, which could not adaptively extend to diverse scenarios. This result potentially proves the significance of our *object-lane interaction weighting*, which enables the model to adaptively disable the object-lane interaction feature when object importance weakly rely on *object-lane interaction* (*e.g.*, static cars on the roadside).

To further analyze *object-lane interaction weighting*, we visualize its output (*i.e.*, p_c in Eq. (13)). Some examples are illustrated in Fig. 3 where objects with blue masks are penalized (*i.e.*, object-lane interaction is disabled, $p_c=\alpha$) and objects with yellow masks are not penalized (*i.e.*, object-lane interaction is enabled, $p_c=1$). In Fig. 3a and Fig. 3b, the *object-lane interaction weighting* penalizes the cars on both sides of the road. The results make sense since the static cars on roadsides are factually not interacting with lanes. In Fig. 3c and Fig. 3d, oncoming cars from the opposite direction and the car on the current lane are not penalized, since these cars are interacting with lanes. We note the yellow mask do not signal the important object. Instead, it indicates that the *object-lane interaction* is enabled.
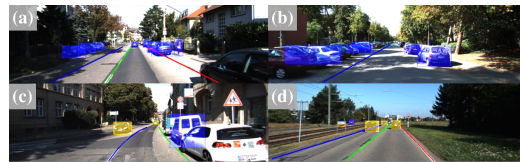

Figure 3: Visualization of *object-lane interaction weighting*.

6   Conclusion

On-road object importance estimation is significant for various applications in the fields of assisted driving and autonomous driving. The dilemmas of current research are two fold: **1)** the scarcity of large-scale publicly available datasets hinder the development of on-road object importance estimation, and **2)** existing methods are relatively simple to handle complex and diverse traffic scenarios. In response to the dilemmas, this paper contributes a new dataset and proposes a model with multi-fold top-down guidance. A large range of experiments demonstrate the superiority of our proposed model. The main conclusion is that building up the model that comprehensively considers multi-fold top-down guidance (*e.g.*, driver intention, semantic context, and traffic rule) and bottom-up feature (*e.g.*, size, distance, and speed) is a promising way to remarkably push forward the study of on-road object importance estimation.

References

[1] Amadori, P.V., Fischer, T., Demiris, Y.: Hammerdrive: A task-aware driving visual attention model. IEEE Transactions on Intelligent Transportation Systems **23**(6), 5573–5585 (2022). https://doi.org/10.1109/TITS.2021.3055120

[2] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

[3] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2018). https://doi.org/10.1109/TPAMI.2017.2699184

[4] Chen, L., Li, Y., Huang, C., Xing, Y., Tian, D., Li, L., Hu, Z., Teng, S., Lv, C., Wang, J., Cao, D., Zheng, N., Wang, F.Y.: Milestones in autonomous driving and intelligent vehicles—part i: control, computing system design, communication, hd map, testing, and human behaviors. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(9), 5831–5847 (2023). https://doi.org/10.1109/TSMC.2023.3276218

[5] Chen, Y., Nan, Z., Xiang, T.: Fblnet: Feedback loop network for driver attention prediction. In: IEEE/CVF International Conference on Computer Vision. pp. 13325–13334 (2023)

[6] Deng, T., Yang, K., Li, Y., Yan, H.: Where does the driver look? top-down-based saliency detection in a traffic driving environment. IEEE Transactions on Intelligent Transportation Systems **17**(7), 2051–2062 (2016). https://doi.org/10.1109/TITS.2016.2535402

[7] Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H.: Dada: Driver attention prediction in driving accident scenarios. IEEE Transactions on Intelligent Transportation Systems **23**(6), 4959–4971 (2022). https://doi.org/10.1109/TITS.2020.3044678

[8] Gao, M., Tawari, A., Martin, S.: Goal-oriented object importance estimation in on-road driving videos. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 5509–5515 (2019). https://doi.org/10.1109/ICRA.2019.8793970

[9] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012). https://doi.org/10.1109/CVPR.2012.6248074

[10] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169

[11] Guo, J., Kurup, U., Shah, M.: Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. IEEE Transactions on Intelligent Transportation Systems **21**(8), 3135–3151 (2020). https://doi.org/10.1109/TITS.2019.2926042

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[13] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[14] Hong, F.T., Li, W.H., Zheng, W.S.: Learning to detect important people in unlabelled images for semi-supervised important people detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4145–4153 (2020). https://doi.org/10.1109/CVPR42600.2020.00420

[15] Hu, Z., Lv, C., Hang, P., Huang, C., Xing, Y.: Data-driven estimation of driver attention using calibration-free eye gaze and scene features. IEEE Transactions on Industrial Electronics **69**(2), 1800–1808 (2022). https://doi.org/10.1109/TIE.2021.3057033

[16] Hu, Z., Zhang, Y., Li, Q., Lv, C.: A novel heterogeneous network for modeling driver attention with multi-level visual content. IEEE Transactions on Intelligent Transportation Systems **23**(12), 24343–24354 (2022). https://doi.org/10.1109/TITS.2022.3208004

[17] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998). https://doi.org/10.1109/34.730558

[18] Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing **26**(9), 4446–4456 (2017). https://doi.org/10.1109/TIP.2017.2710620

[19] Kung, C.H., Yang, C.C., Pao, P.Y., Lu, S.W., Chen, P.L., Lu, H.C., Chen, Y.T.: Riskbench: A scenario-based benchmark for risk identification. arXiv preprint arXiv:2312.01659 (2023)

[20] Li, C., Chan, S.H., Chen, Y.T.: Droid: Driver-centric risk object identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13683–13698 (2023). https://doi.org/10.1109/TPAMI.2023.3294305

[21] Li, J., Gang, H., Ma, H., Tomizuka, M., Choi, C.: Important object identification with semi-supervised learning for autonomous driving. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 2913–2919 (2022). https://doi.org/10.1109/ICRA46639.2022.9812234

[22] Li, J., Zhang, D., Meng, B., Chen, R., Tang, J., Wang, Y.: Enhancement of target feature regions and intention-driven visual attention selection in traffic scenes. In: Proceedings of the IEEE Intelligent Vehicles Symposium. pp. 404–410 (2022). https://doi.org/10.1109/IV51971.2022.9827298

[23] Li, L., Han, J., Zhang, N., Liu, N., Khan, S., Cholakkal, H., Anwer, R.M., Khan, F.S.: Discriminative co-saliency and background mining transformer for co-salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7247–7256 (2023). https://doi.org/10.1109/CVPR52729.2023.00700

[24] Li, W.H., Hong, F.T., Zheng, W.S.: Learning to learn relation for important people detection in still images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4998–5006 (2019). https://doi.org/10.1109/CVPR.2019.00514

[25] Liu, Y., Zhang, J., Li, Y., Hansen, P., Wang, J.: Human-computer collaborative interaction design of intelligent vehicle—a case study of hmi of adaptive cruise control. In: Proceedings of the International Conference on Human-Computer Interaction. pp. 296–314 (2021)

[26] Malla, S., Choi, C., Dwivedi, I., Hee Choi, J., Li, J.: Drama: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1043–1052 (2023). https://doi.org/10.1109/WACV56688.2023.00110

[27] Nan, Z., Jiang, J., Gao, X., Zhou, S., Zuo, W., Wei, P., Zheng, N.: Predicting task-driven attention via integrating bottom-up stimulus and top-down guidance. IEEE Transactions on Image Processing **30**, 8293–8305 (2021)

[28] Nan, Z., Shu, T., Gong, R., Wang, S., Wei, P., Zhu, S.C., Zheng, N.: Learning to infer human attention in daily activities. Pattern Recognition **103**, 107314 (2020)

[29] Nan, Z., Xiang, T.: Third-person view attention prediction in natural scenarios with weak information dependency and human-scene interaction mechanism. IEEE Transactions on Circuits and Systems for Video Technology **34**(8), 6762–6773 (2024)

[30] Nitta, Y., Isogawa, M., Yonetani, R., Sugimoto, M.: Importance rank-learning of objects in urban scenes for assisting visually impaired people. IEEE Access **11**, 62932–62941 (2023). https://doi.org/10.1109/ACCESS.2023.3287147

[31] Niu, Y., Ding, M., Zhang, Y., Ohtani, K., Takeda, K.: Auditory and visual warning information generation of the risk object in driving scenes based on weakly supervised learning. In: Proceedings of the IEEE Intelligent Vehicles Symposium. pp. 1572–1577 (2022). https://doi.org/10.1109/IV51971.2022.9827382

[32] Ohn-Bar, E., Trivedi, M.M.: Looking at humans in the age of self-driving and highly automated vehicles. IEEE Transactions on Intelligent Vehicles **1**(1), 90–104 (2016). https://doi.org/10.1109/TIV.2016.2571067

[33] Ohn-Bar, E., Trivedi, M.M.: Are all objects equal? deep spatio-temporal importance prediction in driving videos. Pattern Recognition **64**, 425–436 (2017)

[34] Organization, W.H.: World health statistics 2023 (2023)

[35] Pal, A., Mondal, S., Christensen, H.I.: "looking at the right stuff" – guided semantic-gaze for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11880–11889 (2020). https://doi.org/10.1109/CVPR42600.2020.01190

[36] Qiu, Y.: Human-centered hmi design for level 3 automated driving takeover process. In: Proceedings of the IEEE International Conference on Intelligent Computing and Human-Computer Interaction. pp. 43–54 (2023). https://doi.org/10.1109/ICHCI58871.2023.10277815

[37] Ramanishka, V., Chen, Y.T., Misu, T., Saenko, K.: Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

[38] Sharma, N., Garg, R.D.: Real-time iot-based connected vehicle infrastructure for intelligent transportation safety. IEEE Transactions on Intelligent Transportation Systems **24**(8), 8339–8347 (2023). https://doi.org/10.1109/TITS.2023.3263271

[39] Sidorenko, G., Thunberg, J., Sjöberg, K., Fedorov, A., Vinel, A.: Safety of automatic emergency braking in platooning. IEEE Transactions on Vehicular Technology **71**(3), 2319–2332 (2022). https://doi.org/10.1109/TVT.2021.3138939

[40] Singh, G., Akrigg, S., Maio, M.D., Fontana, V., Alitappeh, R.J., Khan, S., Saha, S., Jeddisaravi, K., Yousefi, F., Culley, J., Nicholson, T., Omokeowa, J., Grazioso, S., Bradley, A., Gironimo, G.D., Cuzzolin, F.: Road: The road event awareness dataset for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1036–1054 (2023). https://doi.org/10.1109/TPAMI.2022.3150906

[41] Suzuki, T., Kataoka, H., Aoki, Y., Satoh, Y.: Anticipating traffic accidents with adaptive loss and large-scale incident db. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3521–3529 (2018). https://doi.org/10.1109/CVPR.2018.00371

[42] Tang, X., Yu, J., Su, Y.: Modeling driver's visual fixation behavior using white-box representations. IEEE Transactions on Intelligent Transportation Systems **23**(9), 15434–15449 (2022). https://doi.org/10.1109/TITS.2022.3140759

[43] Tian, H., Deng, T., Yan, H.: Driving as well as on a sunny day? predicting driver's fixation in rainy weather conditions via a dual-branch visual model. IEEE/CAA Journal of Automatica Sinica **9**(7), 1335–1338 (2022). https://doi.org/10.1109/JAS.2022.105716

[44] Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)

[45] Wan, J., Li, X., Dai, H.N., Kusiak, A., Martínez-García, M., Li, D.: Artificial-intelligence-driven customized manufacturing factory: Key technologies, applications, and challenges. Proceedings of the IEEE **109**(4), 377–398 (2021). https://doi.org/10.1109/JPROC.2020.3034808

[46] Wu, T., Sachdeva, E., Akash, K., Wu, X., Misu, T., Ortiz, J.: Toward an adaptive situational awareness support system for urban driving. In: 2022 IEEE Intelligent Vehicles Symposium. pp. 1073–1080 (2022)

[47] Xie, C., Xia, C., Ma, M., Zhao, Z., Chen, X., Li, J.: Pyramid grafting network for one-stage high resolution saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11707–11716 (2022). https://doi.org/10.1109/CVPR52688.2022.01142

[48] Yu, S.Y., Malawade, A.V., Muthirayan, D., Khargonekar, P.P., Faruque, M.A.A.: Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. IEEE Transactions on Intelligent Transportation Systems **23**(7), 7941–7951 (2022). https://doi.org/10.1109/TITS.2021.3074854

[49] Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE Access **8**, 58443–58469 (2020). https://doi.org/10.1109/ACCESS.2020.2983149

[50] Zhang, Z., Tawari, A., Martin, S., Crandall, D.: Interaction graphs for object importance estimation in on-road driving videos. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 8920–8927 (2020). https://doi.org/10.1109/ICRA40945.2020.9197104

[51] Zheng, T., Huang, Y., Liu, Y., Tang, W., Yang, Z., Cai, D., He, X.: Clrnet: Cross layer refinement network for lane detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 888–897 (2022). https://doi.org/10.1109/CVPR52688.2022.00097

[52] Zhou, H., Qiao, B., Yang, L., Lai, J., Xie, X.: Texture-guided saliency distilling for unsupervised salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7257–7267 (2023). https://doi.org/10.1109/CVPR52729.2023.00701

Appendix

This appendix provides specialized terms explanation, additional experimental results, limitations, and experimental details, which are organized as follows:

- § A Explanation of Specialized Terms;
- § B Additional Experimental Results;
- § C Limitations;
- § D Experimental Details.

A Explanation of Specialized Terms

1) Bottom-up feature: the low-level information extracted directly from the input images or video frames using backbone networks.

2) Top-down guidance: the high-level information including semantic understanding, prior knowledge, specific goals, etc.

3) Ego-car: the car capturing video sequences that are used as the input of the model.

4) Intention driving path: the path from the ego-car current position to the intention destination.

5) Intention behaviors: the actions that the driver intends to perform based on their goals (*e.g.*, turning left, going straight, and turning right).

B Additional Experimental Results
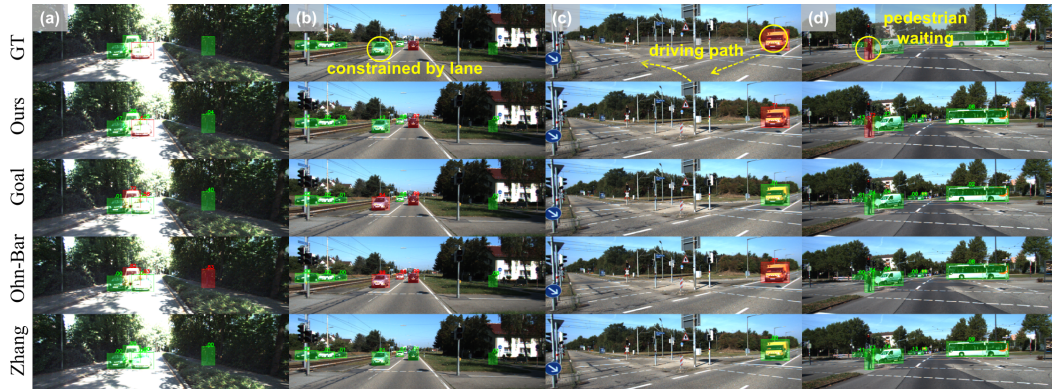
B.1 Qualitative Comparison

Figure 4: Qualitative comparison with baselines (*i.e.*, Goal (8), Ohn-Bar (33), and Zhang (50)). Red boxes represent important objects and green boxes denote unimportant objects.

Four scenarios (a)-(d) are illustrated in Fig. 4. In regular scenarios such as Fig. 4a, most methods function well. However, in complex scenarios, our method exhibits significant superiority. For example, in Fig. 4b, the white car on the opposite lane poses no immediate threat since it should obey the traffic rule to not drive across the solid lane marking. Baselines falsely predict it as an important object as they neglect the influence of traffic rule on object importance, while our method provides the correct estimation by considering the object-lane interaction relation in **TRG**. In Fig. 4c, the ego-car is turning left, and the vehicle on the right side is important because its driving path will risky collide with the intention driving path of ego-car. Our method successfully predicts the important object under the guidance of driver intention, while other methods fail. In Fig. 4d, a pedestrian is waiting to cross the road. Her intention walking path collides with ego-car's intention driving path, thus the pedestrian is important. Our model correctly classifies her as important thanks to the consideration of bottom-up feature and top-down guidance.

B.2 Ablation Study on Object Feature Extraction (OFE)

To validate the rationality of **OFE**, we conduct three experiments, **#1:** only object spatial feature is used; **#2:** only object temporal feature is used; **#3:** both object features are used. The results in Tab. 6 indicate that both spatial and temporal feature of objects serve as valid foundations for accurately evaluating object importance. Removing either of them will lead to the performance decreasing, validating the reasonableness of **OFE** module. The reasons are self-evident, traffic scenarios are highly dynamic and diverse, thus object temporal feature, which reflects motions and behaviors, is significant for object importance estimation. At the same time, object spatial feature conveys rich information such as size, distance, and orientation, thus it is also significant for object importance estimation.

Table 6: Ablation study on OFE.

Method	Spat.	Temp.	AP↑	F1↑
#1	✓		34	35
#2		✓	14	15
#3	✓	✓	**60**	**54**

B.3 Hyperparameter Selection

We perform the experiments on hyperparameter selection, and the results are reported in Tab. 7. In the hyperparameter selection for parameters a and b, we choose the optimal values, $a = 1, b = 1.5$, as the hyperparameters for our model. In the experiments for selecting the hyperparameter α, it is observed that the impact of α is minimal across the three tested values, indicating that our method is robust.

Table 7: Ablation studies on hyperparameter selection.

Parameter	AP↑	F1↑	Parameter	AP↑	F1↑
$a = 1, b = 2.5$	51	41	$\alpha = 0.1$	57	45
$a = 1, b = 2$	56	51	$\alpha = 0.01$	55	52
$a = 1, b = 1.5$	60	54	$\alpha = 0.001$	60	54
$a = 1, b = 1$	49	48			

C Limitations

Figure 5: Failure examples. Top row is GT and bottom row is object importance estimation.

Poor lane marking detection results will limit the performance of our model. In the experiments, we find two kinds of failures. First, as shown in Fig. 5a, lane markings of current lane are falsely detected, thus the white car on the right side is supposed to locate in front of ego-car, leading to the false estimation for the white car. Second, as shown in Fig. 5b, the ego-car moves at a slow speed and lane markings are no detected, thus the model estimates the car in front of the ego-car as an unimportant parking car. Additionally, considering the effect of lane markings on on-road object importance is a small but important step towards modeling the traffic rule for this task. In the future, more traffic rule needs to be considered.

Currently, our model only considers the effect of three types of driver intentions (*i.e.*, turning left, turning right, going straight) on object importance estimation. However, real-world driving scenarios are much more complex. In future work, fine-grained intentions need to be modeled.

D   Experimental Details

D.1   Model Structure Details

Specific details of our model components are introduced in Tab. 8.

Table 8: **Network architecture of our model.** For Region of Interest pooling layer (ROI pooling), we list the output shape. For average pooling (Avg pooling), we list pooling scale. For Long Short-Term Memory network (LSTM), we list input and output dimension and layer number. For multi-head self-attention layer (MHSA), we list the hidden size and the head number. For multi-head cross-attention layer (MHCA), we list the hidden size and the head number. For linear layer (Linear), we list the input and output dimension. For Layer Normalization layer (LN), we list the channel dimension. For multi-layer perceptron network (MLP), we list the input channel dimension and each hidden channel dimension. Note that we use different background colors in the table to distinguish between different modules in our model, where orange represents the **OFE** module, blue represents the **DISG** module, green represents the **TRG** module, and gray represents the **OIE** module.

Layer	Details
1-17	ResNet18(The last FC layer is removed)
18-34	ResNet18(The last FC layer is removed)
35	ROI pooling(10)
36	Avg pooling(16)
37	ROI pooling(10)
38	Avg pooling(16)
39-40	LSTM($512\times10\times10$, 256, 2)
41	MHSA(1024, 8)
42-43	LSTM($512\times10\times10$, 256, 2)
44	MHSA(512, 8)
45	Linear(512, 256), LN(256), ReLU
46	MHCA(1024, 8)
47	Linear($1024\times10\times10$, 256), LN(256), ReLU
48	Linear(20×4, 256), LN(256), ReLU
49	MHCA(256, 8)
50	MLP(256, {1}), Sigmoid
51-52	MLP(256, {128, 2}), Softmax

D.2   Implementation Details

Before the training and inference stages, we utilize CLRNet (51) model with backbone of ResNet101 and pretrained on the CULane dataset to get the L in Eq. (10), and the G in Eq. (5) are generated by applying DeepLabv3 (3). We use OpenCV libary to get the M in Eq. (2). During the training and inference stages, we resize V and M in Eq. (2), and G in Eq. (5) to 320×320 (*i.e.*, $W\times H$=320×320). We set the W'=10, H'=10, C=512, and C'=256. The a and b in Eq. (6) are set as 1 and 1.5, respectively. Then, the β in Eq. (7) is set as 2.2 since the ego-car undergoes steering when the angular velocity is around $\pm$ 2.2 based on the statistical analysis of the IMU data. In addition, the size of m_l, m_s, and m_r is 10×10 (excluding the channel dimension), which is targeted to align with the size of f_s (Eq. (8)). The length of video clip is set as 16. We use SDG optimizer with a weight decay of $5\mathrm{e}^{-4}$ and a momentum of 0.9. We set the batch size as 8, and use the cosine learning strategy with an initial learning rate of $1\mathrm{e}^{-4}$. Our model implementation is based on PyTorch, and experiments are conducted using an NVIDIA RTX3090 GPU.

D.3   Object Bounding Boxes Are Assumed Known

We assume the ground truth of object bounding boxes are given. The purpose is to focus on the on-road object importance estimation while minimizing the influencing factors from the upstream object detection task. This setup is consistent with the existing on-road object importance estimation method (21). Additionally, this setup is reasonable since current object detection models are highly advanced and capable of detecting most objects on the road.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution 1) we claimed is reflected in § 3; The contribution 2) we claimed is reflected in § 5.2; The contribution 3) we claimed is reflected in § 5.2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We dicuss the limitations of our work in § C.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We clearly describe our model in § 4, and the model details can be found in § D.1. The dataset we proposed is fully described in § 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: This paper contributes a new dataset and a model, the dataset will be publicly released in `https://github.com/CQU-ADHRI-Lab/TOI`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental details is specified in § D.2 and § B.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources we used are provided in § D.2.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research we conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our research on-road object importance estimation focuses only on improving the accuracy of importance estimation for road safety purposes. As such, it does not have direct societal impacts beyond the scope of enhancing these technical capabilities.

    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: We are unable to find the license for the dataset and code we used, but they are all open source, and we cite them in the original paper correctly.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The dataset we proposed is well documented and is the documentation provided alongside the assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.