

# OD-RASE: Ontology-Driven Risk Assessment and Safety Enhancement for Autonomous Driving

<sup>†</sup>Kota Shimomura<sup>1,2</sup>, <sup>†</sup>Masaki Nambata<sup>1,2</sup>, Atsuya Ishikawa<sup>3</sup>, Ryota Mimura<sup>3</sup>,  
Koki Inoue<sup>2</sup>, Takayoshi Yamashita<sup>1</sup>, <sup>‡</sup>Takayuki Kawabuchi<sup>3</sup>

<sup>†</sup>Equal Contribution, <sup>‡</sup>Corresponding Author

Chubu University<sup>1</sup>, Elith Inc.<sup>2</sup>, Honda R&D Co., Ltd.<sup>3</sup>

<https://kotashimomura.github.io/odrase/>

## Abstract

Although autonomous driving systems demonstrate high perception performance, they still face limitations when handling rare situations or complex road structures. Such road infrastructures are designed for human drivers, safety improvements are typically introduced only after accidents occur. This reactive approach poses a significant challenge for autonomous systems, which require proactive risk mitigation. To address this issue, we propose OD-RASE, a framework for enhancing the safety of autonomous driving systems by detecting road structures that cause traffic accidents and connecting these findings to infrastructure development. First, we formalize an ontology based on specialized domain knowledge of road traffic systems. In parallel, we generate infrastructure improvement proposals using a large-scale visual language model (LVLM) and use ontology-driven data filtering to enhance their reliability. This process automatically annotates improvement proposals on pre-accident road images, leading to the construction of a new dataset. Furthermore, we introduce the Baseline approach (OD-RASE model), which leverages LVLM and a diffusion model to produce both infrastructure improvement proposals and generated images of the improved road environment. Our experiments demonstrate that ontology-driven data filtering enables highly accurate prediction of accident-causing road structures and the corresponding improvement plans. We believe that this work contributes to the overall safety of traffic environments and marks an important step toward the broader adoption of autonomous driving systems.

## 1. Introduction

State-of-the-art autonomous driving systems achieve highly accurate situational awareness by capturing information beyond what human drivers can process in real time, using

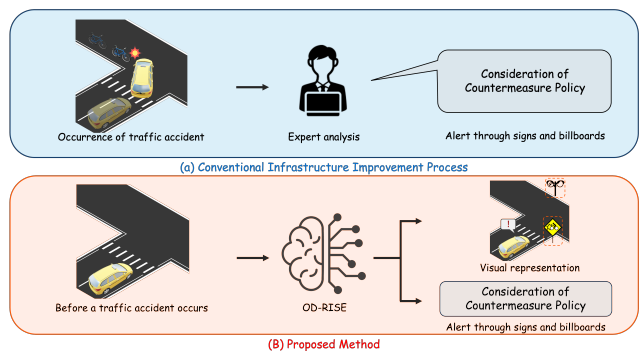


Figure 1. Comparison of various methods for infrastructure improvement design. (a) is based on expert knowledge, while (b) represents our proposed approach. Our method not only outputs infrastructure improvement plans for road structures that cause traffic accidents but also generates visual representations of roads after improvement.

approaches such as BEVPerception methods [34, 50, 56] that efficiently extract 3D features from 2D images, as well as query-based multitask learning methods [19, 46]. These efforts are supported by evaluation platforms that reproduce existing road structures to simulate arbitrary driving scenarios [52, 58] and by benchmark datasets [6, 10, 15, 32, 53]. In addition, data-centric research [26] has focused on generating arbitrary scenes [44, 45, 51] using diffusion models [17] or novel view synthesis [23, 31] to address real-world corner cases. Furthermore, several datasets have been proposed to enhance safe autonomous driving by generating natural language reasoning explanations and evaluating systems ability to understand contextual information [7, 24, 25, 49]. Other methods explicitly learn regions from which vehicles or pedestrians might unexpectedly emerge during driving [39, 55, 57].

These studies substantially contribute to the safety of autonomous driving systems. Nevertheless, the design of road infrastructure has a critical impact on safety, includ-

ing poorly visible intersections, lack of signage, insufficient sidewalk space, or sharp curves. Achieving a higher level of safety requires improving the transportation infrastructure that constitutes the driving environment. As shown in Fig. 1(a), conventional infrastructure improvements are typically carried out after a traffic accident has occurred, with experts analyzing the causes and proposing solutions based on their knowledge. When applied to autonomous driving systems, this process implies that improvements will only be made after an accident caused by an autonomous vehicle, an obviously unacceptable scenario. Consequently, it is essential to expose potential risks in the road environment before an autonomous driving system causes an accident.

In this study, we propose a novel framework that enhances the safety of autonomous driving systems by preemptively identifying road structures that cause traffic accidents and connecting these insights to infrastructure improvements. An overview of our framework is illustrated in Fig. 1(b).

We first construct a dataset that is indispensable for generating infrastructure improvement proposals from images of road structures. To that end, we leverage expert knowledge on road traffic systems to link road structures, their potential risks, and possible improvement methodologies. We formalize this as an ontology based on expert knowledge. We then utilize a large-scale visual language model (LVLM) to generate infrastructure improvement proposals for each road structure. Next, we evaluate these proposals using our ontology, representing both the ontology and the proposals as graphs for graph matching. Only proposals aligning closely with expert knowledge are selected to form our dataset. This ontology-driven filtering enhances the quality and reliability of the dataset. Subsequently, we train the OD-RASE model using the constructed dataset. The OD-RASE model comprises an image encoder [12, 16, 33, 54], a text encoder [9, 28, 54], and a diffusion model. In addition to generating infrastructure improvement proposals for road structure images, the OD-RASE model can visualize the post-improvement road environment. This capability enables even non-experts to visually assess the appropriateness of potential infrastructure modifications.

Our experiments on Mapillary Vistas [32] and BDD100K [53] demonstrate that our method can predict both accident-causing road structures and their corresponding improvement proposals. Moreover, our method shows high robustness in zero-shot prediction on regions outside the training data. These results indicate that our research offers a new perspective on increasing the safety of autonomous driving systems from the standpoint of improving road traffic environments.

Our main contributions can be summarized as follows:

- We propose a novel framework that identifies road structures causing traffic accidents and links them to infrastructure improvements in advance.

- We formalize an ontology that leverages expert knowledge of road traffic systems to represent accident-causing road structures and corresponding infrastructure improvement proposals.
- We improve dataset quality and reliability through ontology-driven data filtering based on expert knowledge.

## 2. Related Work

We first review research on improving road structures to prevent traffic accidents in the domain of road traffic systems. We then survey existing datasets aimed at mitigating traffic risks for autonomous driving.

### 2.1. Road Infrastructure Improvements for Preventing Traffic Accidents

An increasing number of initiatives seek to minimize the damage from accidents by improving road design under the assumption of driver-related human errors [11, 30]. In particular, modifying the road structure itself such as by reducing the number of lanes or installing medians is widely recognized as an effective measure for significantly decreasing the frequency of traffic accidents. In the United States, removing one or more lanes and adding turn lanes has achieved up to a 50% reduction in accidents, and adding a continuous center turn lane can further reduce accidents by up to 65% [1, 22].

Converting intersections into roundabouts is another effective strategy [1, 20, 38]. Roundabouts make it easier to control both the speed and angle of incoming vehicles and have been shown to reduce overall accident rates by around 38%. Additionally, adding left or right turn lanes, reshaping intersections, and installing pavement markings or warning signs have also proven effective [47]. However, these infrastructure improvements primarily focus on human drivers rather than on enhancing the safety of autonomous driving systems. Consequently, there is a need to establish databases specifically designed to inform infrastructure improvements that align with the requirements of autonomous driving.

### 2.2. Datasets for Autonomous Driving

Datasets such as Anticipating Accident [7] and RAMS [35], along with those that provide textual explanations of risks in driving scenarios [24, 25, 49], enable accident prediction and the identification or explanation of high-risk objects. More comprehensive approaches add risk annotations and natural language explanations to driving videos [25, 29]. However, these datasets rely on manual annotations, making large-scale expansion challenging.

To automate the construction of natural language datasets, several approaches have used large-scale visual language models (LVLMs) or large language models (LLMs) for dataset generation [40, 41, 44]. However, verifying the

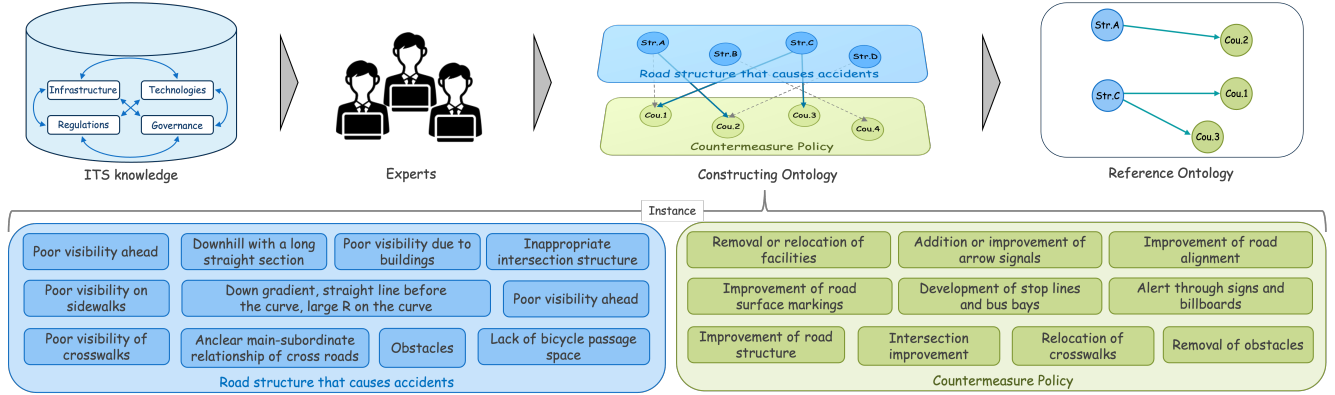


Figure 2. Breakdown of infrastructure improvement process in field of road transportation systems, and overview of how OD-RASE Dataset is constructed on basis of it. Final set of 11 types of road structures causing traffic accidents (top) and 10 types of countermeasures (bottom).

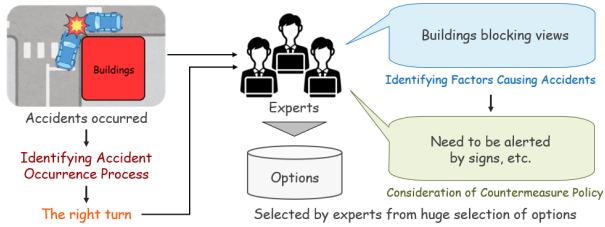


Figure 3. Schematic diagram of the conventional infrastructure improvement process by expert.

quality of generated annotations through human checks incurs a substantial overhead, so typically, only a small, randomly sampled subset undergoes manual review. Xie et al. point out that traditional LVLMs often generate plausible but potentially inaccurate answers by leveraging general knowledge or superficial textual cues [48]. Past work has primarily focused on predicting and describing objects or situations associated with traffic accidents and risks, neglecting the underlying road structures that can fundamentally cause these incidents. Additionally, while automated dataset construction methods reduce the burden of manual annotations, their reliability and quality checks are often limited to a small proportion of the generated data.

### 3. Ontology-Driven OD-RASE Dataset

To train our OD-RASE model, we require a multimodal dataset consisting of images of road environments and corresponding infrastructure improvement proposals. However, no such dataset currently exists, necessitating its creation. Therefore, we attempt to build a dataset by leveraging the conventional process of infrastructure improvements.

#### 3.1. Structuring Infrastructure Improvement Process as Ontology

As shown in Fig. 3, conventional infrastructure improvement processes primarily focus on road environments where a traffic accident has already occurred. First, experts identify the conditions and triggering factors leading to the accident. Next, on the basis of their expert knowledge, they devise an infrastructure improvement process. This process can only be done post-accident. Additionally, because it relies on expert judgment, significant time may be required before an improvement proposal is finalized.

To ensure the safety of autonomous driving systems, it is essential to anticipate potential risks in road environments and perform improvements in advance, even if accidents have not yet occurred. In this research, we focus on summarized information on conventional infrastructure improvement processes [2, 13, 14, 18, 21, 36, 37] and propose a method to structure these processes. This information includes more than 390 cases of road structures, accident conditions, and accident causes, but it lacks a structured format. Consequently, multiple experts in road traffic systems categorize road structures that can cause accidents into 30 types<sup>1</sup>, assigning each structure a corresponding infrastructure improvement proposal. Since some improvements overlap, we reduce the number of distinct proposals to 26 types<sup>2</sup>. We then eliminate any time-dependent factors such as traffic volume or moving vehicles and consolidate similar elements into a single category.

Ultimately, as shown in Fig. 2, the total set of accident-causing road structures is reduced to 11, with infrastructure improvement proposals narrowed down to 10. The combinations of road structures and their respective improvement plans constitute the ontology derived from expert knowledge.

<sup>1</sup>See supplementary materials for details on the 30 types.

<sup>2</sup>See supplementary materials for details on the 26 types.

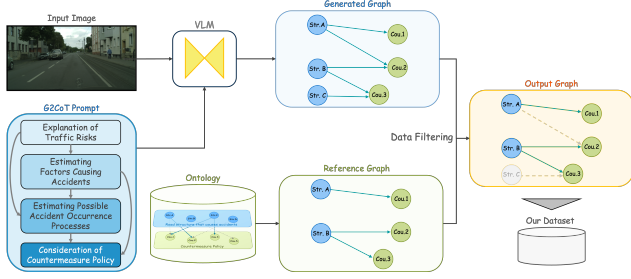


Figure 4. Proposed ontology-driven dataset construction method. Our method allows for fully automatic generation using VLMs and enhances dataset quality and reliability through filtering with reference graphs. Additionally, such filtering further refines dataset’s overall trustworthiness.

### 3.2. Generating Infrastructure Improvement Proposals Based on Expert Reasoning

Fig. 4 illustrates our method for generating candidate infrastructure improvements consistent with expert reasoning. To annotate any arbitrary image of a road structure with an infrastructure improvement plan, we emulate the way experts devise improvements. Experts first consider potential traffic-accident risks in a given image and then infer the conditions under which an accident could occur. Afterward, they predict the specific road structure that triggers the accident. This multi-stage inference process is provided to the VLM as a CoT (chain-of-thought) prompt [43]. For the VLM, we adopt GPT-4o, which has shown high robustness in previous studies [48] and delivers optimal performance on four autonomous driving tasks.

In this workflow, the text generated at each stage is converted into a graph-based prompt [40] and passed to the subsequent stage. We refer to this process as the graph-based grounded CoT prompt (G2CoT)<sup>3</sup>. The infrastructure improvement proposals generated by G2CoT correspond to one of the 10 types described in Sec. 3.1, making it possible to convert them into ontology instances. Similarly, during the G2CoT prompt-generation step, the identified road structures also fall into one of the 11 types described in Sec. 3.1, enabling instance conversion. By leveraging these instantiations of road structures and infrastructure improvements to form our ontology, we can annotate any road-structure image with a proposed improvement plan.

### 3.3. Ontology-Driven Data Filtering Based on Expert Knowledge

We evaluate the validity of the infrastructure improvement proposals generated by GPT-4o using the expert-knowledge ontology introduced in Sec. 3.1. First, we take the ontology based on expert knowledge as a directed reference graph

<sup>3</sup>For more information on this prompt, please refer to the supplementary materials.

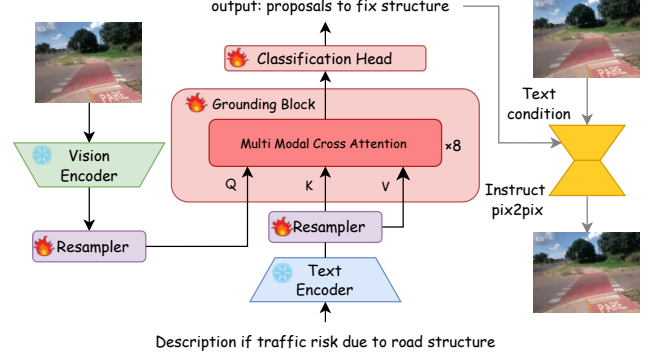


Figure 5. OD-RASE model architecture. Each modality scene images and textual descriptions of traffic risks is encoded, and grounding block captures semantic relationships between image and text.

$G_A = (V_A, E_A)$ . We then treat the generated ontology of improvement proposals as a generated graph  $G_B = (V_B, E_B)$ . We evaluate these two graphs through graph matching. To eliminate elements not present in the reference graph, we compute the common portions of the node sets and edge sets, as in Eq. (1), creating a subgraph  $G'_B = (V'_B, E'_B)$ :

$$V'_B = V_B \cap V_A, \quad E'_B = E_B \cap E_A. \quad (1)$$

This step removes nodes  $(V_B \setminus V_A)$  and edges  $(E_B \setminus E_A)$  that conflict with expert knowledge. However, deleting nodes and edges in this manner may result in isolated nodes with no incoming or outgoing edges. In a directed graph  $G' = (V', E')$ , the set of isolated nodes is defined in Eq. (2):

$$\text{Iso}(G') = \left\{ v \in V' \mid \text{deg}_G^-(v) = 0 \wedge \text{deg}_G^+(v) = 0 \right\}, \quad (2)$$

where  $\text{deg}_G^-(v)$  denotes the in-degree of node  $v$  and  $\text{deg}_G^+(v)$  denotes its out-degree. To remove these isolated nodes, we apply Eq. (3) to  $G'_B = (V'_B, E'_B)$ :

$$V''_B = V'_B \setminus \text{Iso}(G'_B). \quad (3)$$

We then retain only edges whose endpoints belong to  $V''_B$ , yielding the graph  $G''_B$ , which has been filtered in accordance with expert knowledge, as shown in Eq. (4):

$$G''_B = \left( V''_B, E'_B \cap (V''_B \times V''_B) \right). \quad (4)$$

If all edges between any two modules are removed during filtering, we regard such data as untrustworthy and exclude it from the dataset. Through this process, all elements contradicting expert knowledge are discarded, and the final annotation data reflect the same line of reasoning as an expert, thereby achieving high quality.

## 4. OD-RASE Baseline

The aim of this study is to enhance the safety of autonomous driving systems by highlighting potential risks in road structures that lead to traffic accidents. To this end, we propose an architecture capable of predicting an improvement strategy for any road structure. As shown in Fig. 5, our proposed model not only forecasts infrastructure improvement plans but also uses a diffusion model to generate post-improvement road structures. This capability allows non-experts to visually assess the validity of the proposed solutions.

For novice engineers or non-experts, envisioning precise infrastructure improvements can be arduous. Introducing a diffusion model enhances interpretability, thereby facilitating communication among engineers and supporting decision-making.

### 4.1. Multi-Modal Model for Long Contexts

As shown in Fig. 5, our infrastructure improvement proposal model is composed of a vision encoder, a text encoder, and a grounding block.

**Text Encoder:** Let the batch size be  $B$ , the sequence length be  $S$ , and the embedding dimension be  $d$ . For an input text  $\mathbf{x}^t$ , we use a text encoder and obtain  $E_t \in \mathbb{R}^{B \times S \times d}$ , where  $E_t$  is the latent representation of each text token. A linear projection is then used to produce the final text embedding  $T \in \mathbb{R}^{B \times S \times d}$ .

**Vision Encoder:** Similarly, for an input image  $\mathbf{x}^i$ , we use a vision encoder to obtain  $E^i \in \mathbb{R}^{B \times S' \times d}$ , where  $S'$  denotes the number of visual tokens. A linear projection is then used to produce the final image embedding  $I \in \mathbb{R}^{B \times S \times d}$ .

**Grounding Block:** We treat the image embedding  $I$  as the query and the text embedding  $T$  as both the key and value, and we compute a multi-head cross attention as in Eq. (5):

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \quad (5)$$

The final attention output is averaged across the sequence dimension into a vector  $c$ . We then apply a fully connected layer to  $c$  to predict the infrastructure improvement proposal. Since we have 10 predefined categories, the output is the class probability over these 10 categories.

**Training:** An image of a road structure may have more than one valid improvement proposal. Thus, we formulate the task as multi-label classification. Given an input  $x$  composed of both images and text, along with a target  $y_c \in \{0, 1\}$ , the loss is defined as in Eq. (6):

$$\mathcal{L} = - \sum_{c=1}^C \left[ y_c \log p_c + (1 - y_c) \log(1 - p_c) \right], \quad (6)$$

where  $C$  is the number of classes, and  $p_c \in [0, 1]$  is the class probability obtained by applying a sigmoid function to the model logits.

## 4.2. Diffusion-Based Layout Control

To visually depict post-improvement road structures, we adopt a diffusion model with Instruct Pix2Pix for layout control. This module is intended to improve ease of use and explainability. Although it does not directly contribute to the accuracy of forecasts, it plays an important role in the social implementation as a decision support tool.

The image editing process consists of two steps:

1. **Prompt Generation:** From the OD-RASE output, we create a text prompt describing the problematic elements of the road structure and the proposed improvements (e.g., `output1` and `output2` and ...).
2. **Image Editing:** We feed both the original image and the generated text prompt into Instruct Pix2Pix [5], obtaining a new image with the road structure edited in accordance with the proposed improvements.

Through this approach, we move beyond purely textual suggestions, offering a visually rendered depiction of the improved road structure.

## 5. Experiments

In this section, we present experimental results aimed at addressing three questions: (1) Is it possible to derive infrastructure improvement proposals directly from road structure images? (2) Does utilizing an ontology based on expert knowledge enhance the quality of the dataset for infrastructure improvement proposals? (3) Can the proposed method predict infrastructure improvement plans for unseen road structures?

**Training Details.** Below, we summarize our experimental setup. To enable learning of potential traffic accident risks across diverse regions and driving scenarios, we used the Mapillary Vistas [32] and BDD100K [53] datasets, but the proposed approach can be applied agnostically to other datasets. First, we generated candidate infrastructure improvement proposals for these datasets as described in Sec. 3.2.

Next, we evaluated the candidate proposals using the expert-knowledge ontology described in , resulting in final labels for infrastructure improvements. Note that for certain road structures, the final annotation may indicate no improvements needed. We used ResNet-50 [16] and ViT-Base [12] as vision encoders and RoBERTa-Base [28], Flan-T5-xl [9], Long-CLIP [54] as text encoders. Our OD-RASE model was then trained as a multi-label classification task with a batch size of 16 for 25 epochs. The parameters of both the vision and text encoders were frozen during training. We used Recall, Precision, F1-Score, and Accuracy, which are standard metrics for multi-label classification, for our evaluations.

Vision Encoder	Text Encoder	Mapillary				BDD100K			
		Recall	Precision	F1	Acc	Recall	Precision	F1	Acc
ResNet-50[16]	RoBERTa-Base[28]	58.21	73.54	64.98	37.12	74.78	79.75	77.18	45.94
	Flan-T5-xl[9]	64.96	14.42	23.60	0.00	72.85	17.06	27.64	0.00
	Long-CLIP[54]	57.98	75.54	65.60	37.98	72.69	80.96	76.60	45.68
ViT-B[12]	RoBERTa-Base[28]	59.73	78.35	67.79	40.71	74.94	81.62	78.14	47.98
	Flan-T5-xl[9]	57.26	9.81	16.74	0.00	47.70	8.74	14.78	0.00
	Long-CLIP[54]	59.57	78.14	67.60	40.04	73.41	83.31	78.05	47.79
CLIP[33]	RoBERTa-Base[28]	63.70	77.09	69.76	40.71	74.68	82.29	78.30	48.69
	Flan-T5-xl[9]	9.96	3.20	4.85	0.00	7.42	2.61	3.86	0.00
	Long-CLIP[54]	60.47	77.10	67.78	39.96	73.07	83.19	77.80	47.43
Long-CLIP[54]	RoBERTa-Base[28]	64.54	77.09	70.26	42.14	74.83	83.20	78.79	49.48
	Flan-T5-xl[9]	65.67	14.98	24.39	0.00	73.05	17.49	28.22	0.00
	Long-CLIP[54]	61.04	77.21	68.18	40.21	74.65	82.85	78.54	48.62

Table 1. Quantitative evaluation results for predicting infrastructure improvement proposals for road structures. The best second, third best performances are shown in **First**, **Second**, **Third**, respectively.

Modal	Precision	Recall	F1	Acc
Image	57.42	72.37	64.03	34.50
Text	60.02	79.76	68.50	40.63
Image & Text	64.54	77.09	70.26	42.14

Table 2. Ablation study on which modality (image, text, or both) is most effective for predicting infrastructure improvement proposals.



Figure 6. Qualitative examples of infrastructure improvement proposals for road structures. Our OD-RASE model successfully predicted relevant infrastructure improvements.

## 5.1. Predicting Infrastructure Improvements for Road Structures

We first assessed the capability of our OD-RASE model to predict infrastructure improvement proposals for given road structures. According to Tab. 1, the model using Long-CLIP as the vision encoder and RoBERTa-Base as the text encoder achieved the highest performance on both the Mapillary and BDD100K datasets. In contrast, when using Flan-T5-xl as the text encoder, we observed a high recall but low precision,

Data filtering	Precision	Recall	F1	Acc
	33.59	64.85	44.26	0.00
✓	64.54	77.09	70.26	42.14

Table 3. Ablation study on effectiveness of ontology-driven data filtering. Results indicate that filtering improved overall model performance.

suggesting an overestimation of candidate traffic risks and an increased rate of false positives. We conjecture that this is due to our use of 8-bit quantization on the pretrained weights. Tab. 2 presents an ablation study focusing on different input modalities. The results show that combining images and text yielded the best accuracy when predicting infrastructure improvement proposals for road structures. Finally, we show qualitative examples in Fig. 6.

## 5.2. Effectiveness of Dataset Filtering

We investigated the effectiveness of our ontology-driven dataset filtering method proposed in Sec. 3.1. In this experiment, we used Long-CLIP as the vision encoder and RoBERTa-Base as the text encoder, using Mapillary as our dataset. We trained models with and without expert-knowledge-based ontology filtering applied to the candidate proposals. The target task was to predict the road structure responsible for causing traffic accidents, and the evaluation was conducted on the post-filtered data.

Tab. 6 shows the performance with and without data filtering. When the training data were not filtered, the accuracy on the filtered evaluation set was notably low. Specifically, in the absence of filtering, the model demonstrated a high F1-Score of 44.26 pt but a low Accuracy of 0.00 pt, making it difficult to correctly identify the road structures leading to accidents. In contrast, once data filtering was used, the model achieved an F1-Score of 70.26 pt and an Accuracy of

Method		Recall		Precision		F1-Score		Accuracy	
Vision Encoder	Text Encoder	val	test	val	test	val	test	val	test
<b>Ours Baseline</b>									
ResNet-50[16]	RoBERTa-Base[28]	64.38	65.49	69.08	68.91	66.65	67.16	34.97	34.52
ViT-B[12]		62.49	63.65	74.34	73.18	67.90	68.08	38.02	37.12
CLIP[33]		60.64	63.09	73.76	74.26	66.56	68.22	38.02	37.28
Long-CLIP[54]		62.68	62.34	74.91	75.57	68.25	68.32	39.29	38.96
<b>Generalist Models</b>									
	GPT-4o[3]	43.07	61.04	20.83	24.97	26.49	34.27	19.94	23.08
	LLaVA-1.5[27]	39.78	42.17	22.67	23.71	22.51	23.75	15.88	16.82
	Qwen2-VL[42]	41.68	40.99	17.94	17.93	24.09	23.97	15.46	15.55
	Phi-3[4]	41.70	41.27	22.01	22.13	27.48	27.47	17.47	17.48
	InternVL2[8]	38.23	34.57	14.87	14.18	20.17	18.83	12.73	11.97

Table 4. Quantitative evaluation of our OD-RASE model versus generalist models on infrastructure improvement proposal tasks in zero-shot setting. Model was trained on BDD100K and evaluated on Mapillary.

Method		Recall		Precision		F1-Score		Accuracy	
Vision Encoder	Text Encoder	val	test	val	test	val	test	val	test
<b>Ours Baseline</b>									
ResNet-50[16]	RoBERTa-Base[28]	60.25	59.44	79.45	79.64	68.53	68.08	34.57	33.20
ViT-B[12]		63.00	62.21	85.46	86.12	72.53	72.24	41.00	40.62
CLIP[33]		67.66	66.86	84.51	84.87	75.15	74.80	43.63	43.17
Long-CLIP[54]		70.22	69.17	83.68	84.31	76.36	76.00	45.79	44.76
<b>Generalist Models</b>									
	GPT-4o[3]	54.26	50.73	22.79	21.54	31.19	29.38	21.64	20.40
	LLaVA-1.5[27]	47.21	46.71	26.39	26.91	27.27	27.47	19.50	19.62
	Qwen2-VL[42]	28.00	27.88	14.34	14.47	18.40	18.49	11.64	11.65
	Phi-3[4]	52.15	51.35	21.69	21.60	28.94	28.73	18.38	18.24
	InternVL2[8]	25.69	25.11	12.49	12.21	15.91	15.62	10.01	9.79

Table 5. Quantitative evaluation of our OD-RASE model versus generalist models on infrastructure improvement proposal tasks in zero-shot setting. Model was trained on Mapillary and evaluated on BDD100K.

42.14 pt, indicating a more robust learning outcome. From this experiment, we conclude that our proposed ontology-driven filtering, leveraging expert knowledge, is highly effective for exposing the potential risks of road structures that may trigger traffic accidents.

### 5.3. Zero-Shot Prediction

We next evaluated whether the proposed OD-RASE model can predict infrastructure improvement proposals for previously unseen road structures, using a zero-shot setting. We also examined whether state-of-the-art large-scale visual language models (LVLMs), referred to as generalist models, such as GPT-4o [3], LLaVA-1.5 [27], and Qwen2-VL [42], have any general knowledge of infrastructure improvements.

Tab. 4 shows the results for models trained on BDD100K and evaluated on the validation and test sets of Mapillary. Among our baseline variants, the combination of Long-CLIP as vision encoder and RoBERTa-Base as text encoder yielded an F1-Score of 68.32 pt and an Accuracy of 38.96 pt on the Mapillary test set. In contrast, using a generalist

model like Phi-3 yielded an F1-Score of 27.47 pt and an Accuracy of 17.48 pt, indicating difficulty in predicting improvement proposals for an unseen domain.

Tab. 5 shows the results for models trained on Mapillary and evaluated on the validation and test sets of BDD100K, confirming similar trends. These results highlight the necessity of expert-based ontology-driven data filtering.

Indeed, our experiments on two distinct datasets and tasks confirm that existing generalist models lack sufficient domain knowledge to identify accident-causing road structures and propose suitable infrastructure improvements. Simply put, it is difficult to rely solely on these generalist models to detect and fix roads that may lead to traffic accidents.

### 5.4. Diffusion-Based Layout Control

Finally, we show examples of images edited using Instruct Pix2Pix [5] based on the infrastructure development proposals predicted by the OD-RASE model in Fig. 7 and the quantitative evaluation results in Tab. 6. As no ground-truth images for post-improvement roads were available, we pro-



Figure 7. Examples of layout control using *Instruct Pix2Pix* [5], based on OD-RASE-predicted infrastructure improvement plans.

Model	FID↓	Prompt Faithfulness		
		None	Partial	Full
Instruct Pix2Pix	8.5	23.02	22.75	54.23

Table 6. Quantitative evaluation results of the generated images.

vide a qualitative assessment and a quantitative assessment based on FID and expert human evaluation. Experts judged each generated image as “Full”, “Partial”, or “None” based on prompt Faithfulness.

Tab. 6 shows that the images generated by our method are highly interpretable and serve as an effective decision support tool for experts. In the first column, where the proposal was “Improvement to road alignment,” the branching of the road was made more visible from a distance. In scenarios where the proposals were “Improvement to road alignment, improvement to road structure, and alert through signs and billboards” or “Alert through signs and billboards and improvement to road alignment,” the road changes to two lanes, and new signs were visibly added.

Even when multiple proposals are combined, our method allows non-experts to intuitively understand the modifications through visual presentation. For “Improvement of road surface markings,” the white lane markers appear brighter and more pronounced, demonstrating the ability to generate images reflecting proposed improvements. Such visualizations can aid stakeholders in constructing road environments that enhance the safety of autonomous driving systems.

## 6. Limitation

Our current study has several limitations. We categorized road structure improvement proposals based on expert knowledge, excluding those requiring time-series analysis. Furthermore, our analysis relies solely on front-view images

from onboard cameras. Future work could incorporate video inputs or comprehensive road-structure data (e.g., road gradients, intersection curvatures, GIS data) to suggest more sophisticated improvements.

Quantitatively assessing the effectiveness and accident reduction rates of our proposed improvements is challenging, as it would require a traffic simulator capable of editing road structures. Additionally, reflecting the importance and urgency of infrastructure improvement plans quantitatively remains difficult. However, the model’s prediction probability can serve as a confidence level, and incorporating external data like regional traffic volumes and accident rates could help account for importance and urgency more explicitly.

## 7. Conclusion

Although various studies have consolidated the infrastructure improvement process into references [2, 13, 14, 18, 21, 36, 37], these resources are neither structured nor applicable as datasets in computer vision contexts. Drawing on expert knowledge of road traffic systems, we structured road configurations, their potential risks, and methodologies for infrastructure development, defining them as an ontology. Using this ontology, we constructed a high-quality multimodal dataset of infrastructure improvement proposals. Experimental results show that by using ontology-driven data filtering, we can accurately predict both the road structures that cause accidents and their corresponding improvement plans. We also found that state-of-the-art LLMs alone are insufficient for exposing such potential road risks. We believe our work paves the way for safer transportation environments and further advancement of autonomous driving systems.

## References

- [1] Road diet case studies: Santa monica, california ocean park boulevard. U.S. Department of Transportation Federal Highway Administration, pages 14–15, 2015. [2](#)
- [2] Road infrastructure guidelines: New eu-wide guidelines to assess safety of road infrastructure. European Commission: Mobility & Transport - Road Safety, 2023. [3](#), [8](#)
- [3] GPT-4o system card, Accessed:2024-08-06. [7](#)
- [4] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024. [7](#)
- [5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. [5](#), [7](#), [8](#)
- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. [1](#)
- [7] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In Proceedings of the Asian Conference on Computer Vision (ACCV). Springer, 2016. [1](#), [2](#)
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. [7](#)
- [9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53, 2024. [2](#), [5](#), [6](#)
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016. [1](#)
- [11] Mathew Varghese Kavi Bhalla Denny John Ashrita Saran Dinesh Mohan, Geetam Tiwari and Howard White. Protocol: Effectiveness of road safety interventions: An evidence and gap map. Campbell Systematic Reviews, 16(1):2–16, 2020. [2](#)
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2021. [2](#), [5](#), [6](#), [7](#)
- [13] Wenlu Du, Ankan Dash, Jing Li, Hua Wei, and Guiling Wang. Safety in traffic management systems: A comprehensive survey. Designs, 7(4), 2023. [3](#), [8](#)
- [14] Dorin-Ion Dumitrascu. Influence of road infrastructure design over the traffic accidents: A simulated case study. Infrastructures, 9(9), 2024. [3](#), [8](#)
- [15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2012. [1](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016. [2](#), [5](#), [6](#), [7](#)
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS), 2020. [1](#)
- [18] Maher Holozadah and Sahar Tawfiq. Road safety audit for the intersection of cedar road and brainard road and i-271 interchange with brainard road and cedar road. Northeast Ohio Areawide Coordinating Agency, 2013. [3](#), [8](#)
- [19] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. [1](#)
- [20] Douglas W. Harwood Ingrid B. Potts and Karen R. Richard. Relationship of lane width to safety for urban and suburban arterials. Transportation Research Board, 2023(1), 2007. [2](#)
- [21] Sam Thompson David Phipps Carlos Moya Shawn A. Troy Mary Bea Kolbe Christopher Oliver Phillip Vereen Dom Ciaramitaro Joe Seymour Tim Williams Kimberly Hinton Jason Schronce Brian Wert Hanna Cockburn Kristina Solberg Ed Johnson Julie Bogle, Sarah Lee and Jimmy Travis. Action plan for implementing pedestrian crossing countermeasures at uncontrolled locations. North Carolina Department of Transportation, 2018. [3](#), [8](#)
- [22] Jennifer Atkinson Thomas Welch Heather Rigdon Richard Retting Stacey Meekins Eric Widstrand Keith Knapp, Brian Chandler and R.J. Porter. Road diet informational guide. U.S. Department of Transportation Federal Highway Administration, 2010. [2](#)
- [23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 42(4), 2023. [1](#)
- [24] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In European Conference on Computer Vision (ECCV), 2018. [1](#), [2](#)
- [25] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for

- self-driving vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1, 2
- [26] Lincan Li, Wei Shao, Wei Dong, Yijun Tian, Qiming Zhang, Kaixiang Yang, and Wenjie Zhang. Data-centric evolution in autonomous driving: A comprehensive survey of big data system, data mining, and closed-loop technologies. arXiv preprint arXiv:2401.12888, 2024. 1
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in Neural Information Processing Systems (NeurIPS), 2023. 7
- [28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2020. 2, 5, 6, 7
- [29] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023. 2
- [30] Mari Svolsbru Blair Turner Matts-Åke Belin, Anders Hartmann and Michael S. Griffith. Public roads - winter 2022. U.S. Department of Transportation Federal Highway Administration, 85(4), 2022. 2
- [31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In European Conference on Computer Vision (ECCV), 2020. 1
- [32] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017. 1, 2, 5
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), 2025. 2, 6, 7
- [34] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 1
- [35] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2
- [36] R.A. Retting, B.N. Persaud, P.E. Garder, and D. Lord. Crash and injury reduction following installation of roundabouts in the United States. American Journal of Public Health, 91(4): 628–631, 2001. 3, 8
- [37] Richard A. Retting, Allan F. Williams, David F. Preusser, and Helen B. Weinstein. Classifying urban crashes for countermeasure development. Accident Analysis Prevention, 27(3): 283–294, 1995. 3, 8
- [38] Per E. Garder Richard A. Retting, Bhagwant N. Persaud and Dominique Lord. Crash and injury reduction following installation of roundabouts in the united states. American Journal of Public Health, 91(4):628–631, 2001. 2
- [39] Kota Shimomura, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Potential risk localization via weak labeling out of blind spot. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024. 1
- [40] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In European Conference on Computer Vision (ECCV), 2024. 2, 4
- [41] Pittawat Taveekitworachai, Pratch Suntichaikul, Chakarida Nukoolkit, and Ruck Thawonmas. Speed up! cost-effective large language model for adas via knowledge distillation. In 2024 IEEE Intelligent Vehicles Symposium (IV), 2024. 2
- [42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 7
- [43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS), 2022. 4
- [44] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1, 2
- [45] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1
- [46] Xinchuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. 2024. 1
- [47] Wilson. 2+1 roads - swedish innovation canadian rural road solution? CARSP/PRI JOINT VIRTUAL CONFERENCE, 2021. 2
- [48] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. arXiv preprint arXiv:2501.04003, 2025. 3, 4
- [49] Y. Xu, X. Yang, L. Gong, H. Lin, T. Wu, Y. Li, and N. Vasconcelos. Explainable object-induced action decision

- for autonomous vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 2
- [50] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Bevformer v2: Adapting modern image backbones to bird’s-eye-view recognition via perspective supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 1
- [51] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1
- [52] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In International Conference on Learning Representations, 2024. 1
- [53] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 2, 5
- [54] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In European Conference on Computer Vision (ECCV), 2024. 2, 5, 6, 7
- [55] Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, and Yun Fu. Oostraj: Out-of-sight trajectory prediction with vision-positioning denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1
- [56] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv:2205.09743, 2022. 1
- [57] Jiacheng Zhou, Masahiro Hirano, and Yuji Yamakawa. High-speed recognition of pedestrians out of blind spot with pre-detection of potentially dangerous regions. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2022. 1
- [58] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1