BEV-TSR: Text-Scene Retrieval in BEV Space for Autonomous Driving

Tao Tang[1*‡], Dafeng Wei[2*], Zhengyu Jia[2*], Tian Gao[2*], Changwei Cai[2], Chengkai Hou[2**], Peng Jia[2], Kun Zhan[2], Haiyang Sun[2], Jingchen Fan[2], Yixing Zhao[2], Xiaodan Liang[1†], Xianpeng Lang[2], Yang Wang[2†]**

[1]Shenzhen Campus of Sun Yat-sen University
[2]Li Auto Inc.
{weidafeng, jiazhengyu, gaotian, wangyang25}@lixiang.com

Abstract

The rapid development of the autonomous driving industry has led to a significant accumulation of autonomous driving data. Consequently, there comes a growing demand for retrieving data to provide specialized optimization. However, directly applying previous image retrieval methods faces several challenges, such as the lack of global feature representation and inadequate text retrieval ability for complex driving scenes. To address these issues, firstly, we propose the **BEV-TSR** framework which leverages descriptive text as an input to retrieve corresponding scenes in the Bird's Eye View (BEV) space. Then to facilitate complex scene retrieval with extensive text descriptions, we employ a large language model (LLM) to extract the semantic features of the text inputs and incorporate knowledge graph embeddings to enhance the semantic richness of the language embedding. To achieve feature alignment between the BEV feature and language embedding, we propose Shared Cross-modal Embedding with a set of shared learnable embeddings to bridge the gap between these two modalities, and employ a caption generation task to further enhance the alignment. Furthermore, there lack of well-formed retrieval datasets for effective evaluation. To this end, we establish a multi-level retrieval dataset, nuScenes-Retrieval, based on the widely adopted nuScenes dataset. Experimental results on the multi-level datasets show that BEV-TSR achieves state-of-the-art performance, e.g., 85.78% and 87.66% top-1 accuracy on scene-to-text and text-to-scene retrieval respectively.

Introduction

The past few years have witnessed rapid advancements in the autonomous driving industry (Ma et al. 2022; Li et al. 2023a), which transitioned from a phase of data scarcity to one where data is abundant, owing to the substantial data generated by both data collection vehicles and crowdsourcing vehicles. However, the mere accumulation of uniformly distributed data is no longer sufficient to achieve significant improvements (Long et al. 2022). For example, if we aim to ensure effective performance on rural roads, it is imperative to retrieve a sufficient amount of rural road data to fine-tune the models. Therefore, data mining has become an essential working schema to provide specialized optimization for autonomous driving models, and a well-designed retrieval method plays a crucial role in the closed-loop data-driven pipeline of autonomous driving data (Fu et al. 2024; Li et al. 2024).

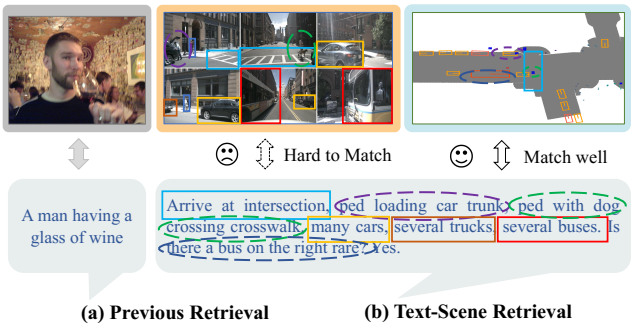

Figure 1: **(a)** Existing methods are primarily tailored for simple retrieval scenarios. **(b)** On the contrary, autonomous driving scenarios are challenging with numerous traffic participants and road elements. Then the BEV space offers a clearer global context of the scene than the previous image space, which aligns well with the textual query and serves as an ideal retrieval space. To this end, we propose the novel BEV-TSR framework for text-scene retrieval in autonomous driving, which retrieves scenes in BEV space and demonstrates a significant capability to retrieve traffic scenarios.

On the other hand, cross-modal image-text retrieval (ITR) has been a longstanding research task and presents a significant advancement over the past few years as to the prosperity of deep models for language and vision (Cao et al. 2022; Ma et al. 2021). However, as shown in Fig. 1, despite significant recent progress in the field of ITR, when applied to challenging autonomous driving scenarios, modern retrieval methods often struggle to achieve satisfactory results. While previous methods (Fang et al. 2023; Li et al. 2023b; Zhai et al. 2023) have been successful in handling simpler tasks such as identifying objects like "a brown dog" or more complex descriptions like "a man having a glass of wine," autonomous driving datasets contain numerous traffic participants and road elements, e.g., "arrive at intersection, ped loading car trunk,

* Equal contribution
‡Work done during an internship at Li Auto Inc.
** Equal contribution
† Co-corresponding author

ped with dog crossing crosswalk, many cars, several buses." These complex traffic scenes present significant challenges for retrieval models in autonomous driving: models need to possess a comprehensive understanding of the global context of traffic scenes and the ability to comprehend complex and lengthy textual inputs. These requirements go beyond the capabilities of existing retrieval methods.

In this paper, we propose BEV-TSR framework to tackle the barriers from the main components of the retrieval pipeline. The traditional ITR paradigm involves two main steps: (1) *extracting* the representation of images and sentences; and (2) *aligning* the cross-model representations with similar semantics. *For feature extraction*, for the scene image representation, we propose to retrieve corresponding scenes in the Bird's Eye View (BEV) space. As shown in Fig. 1 (b), the image space often presents scattered and truncated elements of traffic scenes, e.g., the two images highlighted within the red box only display partial views of the bus. While the BEV space offers a clearer global context of the scene, which aligns well with the textual query and serves as an ideal retrieval space. For text sentence representation, to facilitate complex scene retrieval with extensive text descriptions, we utilize a large language model (LLM) to extract semantic features from textual inputs. Moreover, we incorporate learned knowledge graph embeddings of autonomous driving which further enhances the understanding and comprehension of the textual elements. Then, *for feature alignment*, with the BEV feature and language embedding in different feature spaces, we propose the Shared Cross-modal Embedding, which utilizes a set of shared learnable embeddings to bridge the gap between the two modalities. Additionally, we employ a caption generation task to further enhance the alignment. Furthermore, there lacks well-structured retrieval datasets for effective evaluation. To address this, we establish a multi-level retrieval dataset, nuScenes-Retrieval, based on the widely adopted nuScenes dataset. Experimental results on the multi-level nuScenes-Retrieval demonstrate that BEV-TSR achieves significant advancements, e.g., 85.78% and 87.66% top-1 accuracy on scene-to-text and text-to-scene retrieval respectively.

Our main contributions can be summarized as follows:

1. We propose the novel BEV-TSR framework for text-scene retrieval in autonomous driving, which retrieves scenes in BEV space and demonstrates a significant capability to understand the global context and retrieve complex traffic scenarios.
2. We leverage an LLM and incorporate knowledge graph embeddings to comprehensively understand complex textual descriptions, offering a higher level of semantic richness in language embedding.
3. We propose Shared Cross-modal Embedding with a set of shared learnable embeddings to align the cross-model features, and employ a caption generation task to further enhance the alignment.
4. We establish a multi-level retrieval dataset, nuScenes-Retrieval, based on the nuScenes dataset, on which our BEV-TSR achieves state-of-the-art performance, e.g., 85.78% and 87.66% top-1 accuracy on scene-to-text and text-to-scene retrieval respectively.

Related Work

Image-Text Retrieval. Image-Text Retrieval (ITR) is a fundamental cross-modal task in computer vision, whose main challenge lies in learning a shared representation of images and texts and accurately measuring their similarity (Cao et al. 2022). Traditional methods for text-to-image retrieval have typically relied on convolutional neural networks (CNNs) as independent encoders to produce representations to measure the similarities between images and textual content (Dong et al. 2014; Noh et al. 2017; Radenović, Tolias, and Chum 2018). Recent years have witnessed a surge in the adoption of transform-based models and large-scale language-image pre-training (Radford et al. 2021; Fang et al. 2023; Li et al. 2023b; Zhai et al. 2023), which achieve state-of-the-art performance across various text-to-image benchmark tasks. However, existing methods are primarily tailored for simple retrieval scenarios with simple text inputs, focusing on single-image retrieval, and are evaluated on datasets such as MSCOCO (Lin et al. 2014) and Flickr30k (Plummer et al. 2015). The potential applications to large-scale complex scenarios remain largely unexplored. In this work, we first delve into the text-scene retrieval of autonomous driving. We find that modern retrieval methods often struggle to achieve satisfactory results when applied to challenging autonomous driving scenarios, as they cannot comprehensively understand the global context of traffic scenes and comprehend complex and lengthy textual inputs. To this end, we propose the BEV-TSR framework which incorporates several elaborate modules for feature extraction and alignment, to achieve accurate text-scene retrieval of autonomous driving.

BEV Space Learning. In recent years, learning powerful representations in Bird's Eye View (BEV) space has been a growing trend and garnered significant attention from both industry and academia (Ma et al. 2022; Li et al. 2023a). BEV approaches have gained popularity in various aspects of autonomous driving, including detection (Wang et al. 2023a; Park et al. 2022), segmentation (Liang et al. 2022; Liu et al. 2023b; Chen et al. 2022), tracking (Lin et al. 2023), and occupancy forecasting (Zhang, Zhu, and Du 2023; Wang et al. 2023b; Tong et al. 2023). The main challenge in learning within the BEV space is the reconstruction of BEV features from the perspective view. Following LSS (Philion and Fidler 2020), some methods like BEVDet series (Huang et al. 2021; Li et al. 2023d,c) predict a distribution over depth bins and lift weighted image features onto BEV. Another branch of work follows DETR3D (Wang et al. 2022) and adopts BEV queries and projects them to get image features. BEVFormer (Li et al. 2022) introduces sequential temporal modeling into multi-view 3D object detection and applies temporal self-attention. PETR (Liu et al. 2022) proposes 3D position embedding in the global system and then conducts global cross-attention to update the queries. These methods have gained popularity due to the advantages of BEV space learning, such as providing an intuitive representation of the world and enabling the representation of objects in BEV,

which is most desirable for subsequent modules in planning and control. These characteristics naturally make BEV space an optimal space for text-scene retrieval. Therefore, in this paper, we propose BEV-TSR, the first text-scene retrieval method in the BEV space.

BEV-TSR

In this section, we introduce our proposed BEV-TSR in detail. We first give a brief problem definition and an overview. Then, we subsequently delve into our key contributions on feature extraction and feature alignment.

Preliminaries

Problem definition. In this study, we concentrate on the task of text-scene retrieval on the driving dataset with paired text sentence $\mathcal{T}$ and a scene collection with images $\mathcal{I}$. Given a textual query q, the retrieval model returns a ranked list of scene images L. Here, L_i represents the i-th ranked image in the list. The objective is to retrieve as many relevant images as possible from the top-k ranked scene images $\mathcal{L}_k$.

Overall architecture. As illustrated in Fig. 2, for feature extraction, our BEV-TSR adapt a BEV encoder to extract the visual BEV embedding b_i, and the textual embedding is enriched by incorporating the knowledge graph embeddings and then embedded as t_i from a language encoder. For feature alignment, we leverage a set of shared learnable embeddings to bridge the gap between the two branches' features. The resulting features are aligned with the contrastive loss based on cosine similarity. Moreover, we also employ caption generation as an auxiliary task to enhance the alignment further. During retrieval, we extract embedding t for the given query text and compute the cosine similarity with the BEV vector $B = \{b_1, b_2, \ldots, b_n\}$ of all images in the candidate pool to identify the top-k similar scene images.

Feature Extraction

BEV Encoder. The utilization of BEV space learning in autonomous driving allows for an intuitive representation of the world, making it an optimal choice for text-scene retrieval tasks. To extract BEV features, we employ a dedicated BEV encoder for generating BEV embeddings B from the visual sequence inputs $\mathcal{I}$. It is worth noting that a wide range of visual BEV encoders can be employed for this purpose, such as BEVFormer (Li et al. 2022), BEVDet (Huang et al. 2021). These encoders have demonstrated exceptional performance in similar tasks and are capable of capturing informative visual features in the BEV domain.

Text Encoder. To extract comprehensive semantics information from textual input, we utilize recent popular large language models as our text encoder, such as Llama (Touvron et al. 2023) or GPT-3 (Floridi and Chiriatti 2020), which generates language embeddings T from the text sentence $\mathcal{T}$. As these models (Brown et al. 2020; Chowdhery et al. 2022; Chung et al. 2022; Liu et al. 2023a) have demonstrated great generalization power and common sense reasoning ability, indicating their potential to understand the scenarios in the realm of autonomous driving.

Knowledge Graph Prompting Scene text inputs are usually described using discrete categories. For example, when discussing a traffic scene, discrete categories are used to describe the types of vehicles present (e.g., cars, trucks), their colors (e.g., red, blue), their actions (e.g., stopping, turning), and the scene's attributes (e.g., traffic lights, pedestrian crossings). However, only discrete semantic information is insufficient for text-scene retrieval, as it lacks contextual comprehension and fails to capture relationships and associations between elements. To address this, we propose incorporating knowledge graph embedding that provides associative information as a complement to enhance overall scene understanding.

Knowledge graph. Knowledge Graphs (KGs) are a type of multi-relational graphs that store factual knowledge in the real world. It is typically represented as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{S}\}$, where $\mathcal{E}$ denotes the set of entities (nodes), $\mathcal{R}$ denotes the set of relations (the types of edges) and $\mathcal{S}$ represents the relational facts (edges). Facts observed in $\mathcal{G}$ are stored as a collection of triples: $\mathcal{G} = \{(h, r, t)\}$, where each triple consists of a head entity $h \in \mathcal{E}$, a tail entity $t \in \mathcal{E}$, and a relation $r \in \mathcal{R}$ between them, e.g., $\langle$*scene, includes, car*$\rangle$ as illustrated in Fig. 3 (a).

Knowledge graph embedding. With the large-scale and complicated graph structure of the autonomous driving KG. We then need to learn knowledge graph embeddings (KGEs) to represent the entities and relations in low-dimensional vector space while also maintaining the semantics contained in the original KG. As suggested by Wickramarachchi et al. (Wickramarachchi, Henson, and Sheth 2020), we adopt the semantic transitional distance-based modeling (Bordes et al. 2013). Specifically, we adopt a distance-based scoring function to optimize the embeddings. For each triplet (h, r, t), the scoring function is defined as follows:

$$f_r(h, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p, \tag{1}$$

where $p = 1$ or $p = 2$, and $\mathbf{h}$, $\mathbf{t}$, and $\mathbf{r}$ represent the embedding of the head entity, tail entity, and the relation between entities, respectively. With $\mathbf{r}$ represents a translation vector from $\mathbf{h}$ to $\mathbf{t}$, $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ when the triple (h, r, t) holds true. With the scoring function, we learn KGEs that capture the relationships depicted by the triplets within the autonomous driving KG.

Semantic representation fusion. Then, we utilize the obtained knowledge graph embeddings to enrich the semantic representation of the original text embedding. Practically, as illustrated in Fig. 3 (b), we index the keywords from the graph that appear in the text input and concatenate their knowledge graph embeddings into the text embedding sequence in the order of their occurrence.

Feature Alignment

Shared Cross-modal Embedding Through the BEV encoder and LLM text encoder with KGP, we obtain the BEV feature and language embedding in two different feature spaces. Then directly aligning these cross-modal features leads to unsatisfactory outcomes. To address this challenge,

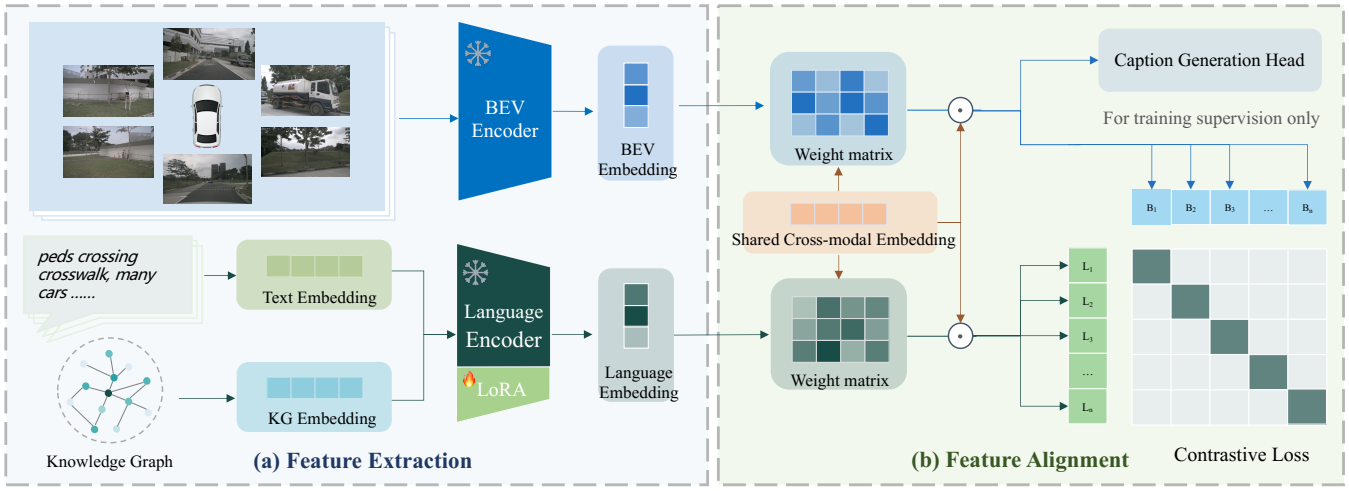

Figure 2: **Overall framework of BEV-TSR. (a) Feature Extraction.** For the visual branch, the BEV encoder extracts the BEV embedding from surrounding camera images. For the textual branch, the text embedding is enriched by incorporating the knowledge graph embedding and then fed into a language encoder to generate language embedding. **(b) Feature Alignment.** First, a set of shared learnable embeddings are employed to bridge the gap between the two branches' features. Then, the resulting features are aligned with the contrastive loss. Moreover, caption generation auxiliary task further enhances the alignment.

we propose the Shared Cross-modal Embedding (SCE) approach, inspired by the Q-Former in BLIP2 (Li et al. 2023b). SCE utilizes a set of shared learnable embeddings to bridge the gap between the two modalities by re-projecting their respective features into a shared feature space, similar to an attention mechanism.

Specifically, the learnable shared cross-modal embeddings are denoted as $C = \{c_1, c_2, \ldots, c_k\}$, and the BEV features can be reshaped and compressed into a sequence of BEV embeddings as $B = \{b_1, b_2, \ldots, b_n\}$. For each embedding c_i from C in the shared cross-modal embeddings sequence, the similarity between each feature b_j in the BEV sequence and the embedding c_i is computed as $s_{ij} = sim(c_i, b_j)$, where sim represents cosine similarity. Then the maximum similarity is obtained between the BEV embeddings sequence B and the embedding t_i as $r_i = \max_j(s_{ij})$. For cross-modal embeddings C, we have $R = \{r_1, r_2, \ldots, r_k\}$. Then, with the softmax function, R can be converted into weights: $w_i^b = \frac{e^{r_i}}{\sum e^r}$. The re-projected BEV embeddings are obtained by multiplying the weights with the shared cross-modal embeddings: $B' = \{b'_1, b'_2, \ldots, b'_n\} = \{w_1^b c_1, w_2^b c_2, \ldots, w_k^b c_k\}$. Similarly, for the language embeddings T, we obtain $T' = \{w_1^t c_1, w_2^t c_2, \ldots, w_k^t c_k\}$. Subsequently, the obtained BEV and language embeddings B' and L' are aligned with contrastive loss:

$$\mathcal{L}_{SCE} = \mathcal{L}_{\text{cl}}^{\text{t2s}} + \mathcal{L}_{\text{cl}}^{\text{s2t}}, \tag{2}$$

$$\mathcal{L}_{\text{cl}}^{\text{t2s}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\exp(sim(t'_i, b'_i)/\tau)}{\sum_{j=1}^{N}\exp(sim(t'_i, b'_j)/\tau)}\right), \tag{3}$$

$$\mathcal{L}_{\text{cl}}^{\text{s2t}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\exp(sim(b'_i, t'_i)/\tau)}{\sum_{j=1}^{N}\exp(sim(b'_i, t'_j)/\tau)}\right), \tag{4}$$

where N is the batch size and τ is a temperature hyperparameter, and $t2s$ and $s2t$ denotes text-to-scene and scene-to-test respectively.

Practically, the cross-modal embeddings shared between the BEV and text branches serve as a bridge in a shared feature space. It enables maintaining the distinct shapes of the original features while achieving alignment between the BEV and text modalities within a unified embedding space.

Caption Generation To further enhance the alignment between the BEV embedding and the language embedding, we introduce an auxiliary caption generation task based on the BEV embeddings. We utilize a lightweight transformer-based decoder, with the corresponding text description of the BEV sample serving as the supervision label:

$$\mathcal{L}_{CG} = \text{CrossEntropy}(P_{logits}, T), \tag{5}$$

where P_{logits} is the logits of predictive text tokens and T is the target text tokens. By incorporating this caption generation task, we strengthen the relationship between the BEV embedding and the generated textual description, thereby improving the overall text-scene retrieval performance.

Overall Optimization

To summarize, the overall optimization target of our BEV-TSR is formulated as:

$$\mathcal{L} = \mathcal{L}_{SCE} + \lambda\mathcal{L}_{CG} \tag{6}$$

where $\mathcal{L}_{SCE}$ and $\mathcal{L}_{CG}$ are defined in Eq. (2) and Eq. (5), and λ indicate the weight balance coefficient.

nuScenes-Retrieval Dataset

Although we have developed the BEV-TSR framework for autonomous driving retrieval, the driving community

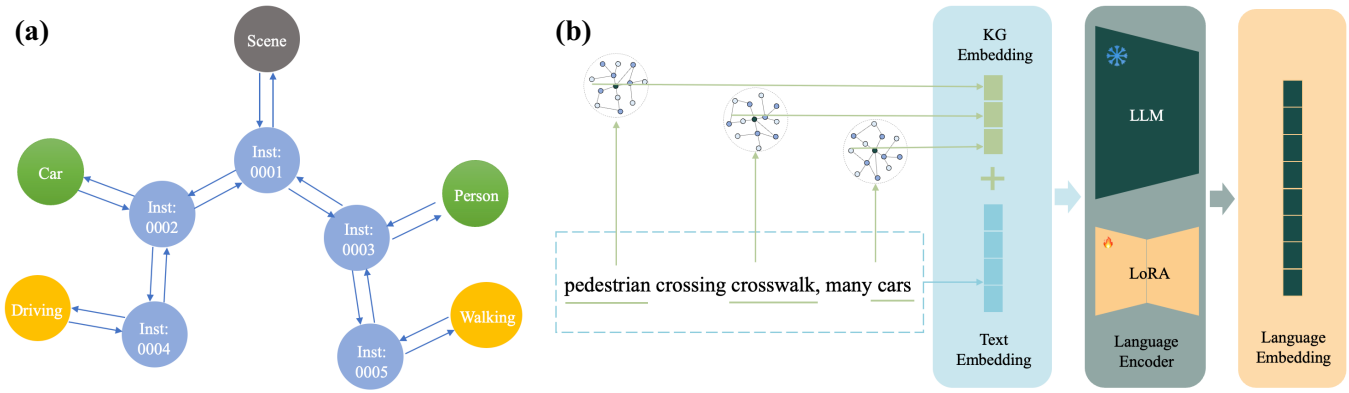

Figure 3: **Knowledge graph prompting.** **(a)** The knowledge graph embeddings are learned from the autonomous driving knowledge graph. Each node in the graph corresponds to a keyword relevant to autonomous driving, and the embeddings associated with these nodes capture the associative representation of autonomous driving keywords. **(b)** Subsequently, these keyword knowledge graph embeddings are concatenated with the text embedding, thereby expanding the semantic representation of the encoded text, and then embedded from a language encoder.

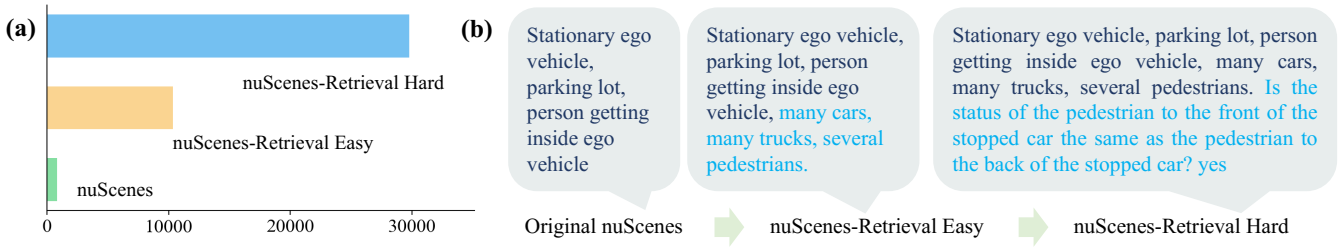

Figure 4: **(a)** The number of text descriptions. **(b)** A case on the nuScenes-Retrieval dataset.

lacks well-structured retrieval datasets for effective evaluation. The only available dataset with scene sensor data and open text descriptions is the nuScenes dataset (Caesar et al. 2020). However, the provided text descriptions in nuScenes are very simplistic and lack detailed scene information, as shown in Fig. 4 (a), the nuScenes dataset consists of over 30,000 samples but only has 848 distinct text sentences available, which leads to significant repetition in the provided text descriptions. To address these limitations, we have further constructed the nuScenes-Retrieval dataset based on the nuScenes dataset, and the toolkit codes are attached in the supplement materials and will be public.

nuScenes-Retrieval Easy. As shown in Fig. 4 (b), to reduce text repetition, we enhance the original captions by supplementing them with perception results. Specifically, for each keyframe sample, we extract obstacle information from the detection labels. Moreover, we quantify the frequency of occurrence for each obstacle type, then render the obstacles with quantity descriptors, such as "many cars, several trucks, one bus". The supplemented text is concatenated after the original caption text, resulting in a more comprehensive scene description for training and evaluation. This modified dataset is referred to as *nuScenes-Retrieval Easy*, which provides more than 10k text descriptions.

nuScenes-Retrieval Hard. To further improve the informative diversity of the textual descriptions, we have noticed research efforts, i.e., NuScenes-QA (Qian et al. 2023), which is a visual question-answer (VQA) dataset generated from the nuScenes dataset, consisting of over 34,000 scenes with more than 460,000 question-answer pairs. These pairs are generated from perception information using designed QA templates and cover various aspects, including object detection, scene classification, ego-vehicle decision-making, and decision reasoning. As shown in Fig. 4 (b), to further enrich the text input with semantic information related to potential decision descriptions, we concatenate the questions and answers together and append them at the end of the previous description. This extended dataset is dubbed *nuScenes-Retrieval Hard* with 29k text descriptions.

Experiment

In this section, we first introduce the experimental setup. Then, we compare BEV-TSR with previous approaches and present the analysis of each component.

Experimental Setup

Datasets and Metric. For the retrieval dataset, we adopt the nuScenes-Retrieval based on the nuScenes dataset (Caesar et al. 2020). We use (R@K, K=1,5,10) as our evaluation metrics of recall accuracy, which is the most commonly used evaluation metric in the retrieval tasks and is the abbreviation for recall at k-th in the ranking list, defined as the proportion of correct matchings in top-k retrieved results.

Method	Retrieve Space	Text Retrieval			Scene Retireval		
		R@1	R@5	R@10	R@1	R@5	R@10
nuScenes-Retrieval Easy							
SigLIP-Base (Zhai et al. 2023)	Front View	0.3683	0.7640	0.8613	0.3924	0.7863	0.8698
	Surrounding View	0.3433	0.7625	0.8593	0.3597	0.7740	0.8672
CLIP-ViT-Base (Radford et al. 2021)	Front View	0.4377	0.8610	0.9569	0.4421	0.9003	0.9795
	Surrounding View	0.4846	0.9085	0.9815	0.4644	0.9258	0.9845
EVA02-Base (Fang et al. 2023)	Front View	0.4919	0.7306	0.7977	0.5585	0.7807	0.8440
	Surrounding View	0.4369	0.7153	0.7986	0.5181	0.7896	0.8637
BEV-TSR (Ours)	BEV Space	**0.8578**	**0.9954**	**0.9994**	**0.8766**	**0.9971**	**0.9997**
nuScenes-Retrieval Hard							
SigLIP-Base (Zhai et al. 2023)	Front View	0.2687	0.6573	0.7661	0.2850	0.6691	0.7670
	Surrounding View	0.2594	0.6487	0.7379	0.2843	0.6501	0.7365
CLIP-ViT-Base (Radford et al. 2021)	Front View	0.2829	0.6501	0.7886	0.2986	0.6888	0.7863
	Surrounding View	0.2904	0.6619	0.7953	0.3163	0.7066	0.7896
EVA02-Base (Fang et al. 2023)	Front View	0.2908	0.6763	0.7860	0.3064	0.6980	0.7936
	Surrounding View	0.2774	0.6395	0.7318	0.2965	0.6538	0.7430
BEV-TSR (Ours)	BEV Space	**0.6608**	**0.9912**	**0.9997**	**0.6790**	**0.9862**	**0.9991**

Table 1: Comparisons with different retrieval methods.

The implementation details are provided in the supplementary material.

Main Results

In Table 1, we compare our BEV-TSR with state-of-the-art retrieval methods, i.e., CLIP-ViT-Base (Radford et al. 2021), SigLIP-Base (Zhai et al. 2023), and EVA02-Base (Fang et al. 2023) on the multi-level nuScenes-Retrieval dataset. To ensure a fair comparison, we extended the previous methods to include six surrounding images. From the results, it can be observed that our BEV-TSR achieves impressive performance (85.78% and 87.66% top-1 accuracy) and outperforms all the other methods by a large margin, both from the front view and six surrounding views. This indicates that our method is effective in retrieving scenes in the BEV space and demonstrates a significant capability to understand the global context and retrieve complex traffic scenarios.

Furthermore, across different levels of the nuScenes-Retrieval dataset, our method consistently achieves superior results. Particularly, a more pronounced improvement is achieved in the Hard level, indicating that our BEV-TSR framework with well-designed modules, is capable of better understanding and retrieving more complex autonomous driving scenes.

Qualitative Results

In Fig. 5, we present the qualitative results on nuScenes-Retrieval dataset. We compare our BEV-TSR with the previous state-of-the-art retrieval method, EVA02 (Fang et al. 2023). EVA02, as an image space retrieval model, fails to find scenes that correspond well to the textual queries. In contrast, our BEV-TSR achieves more precise retrieval results, which retrieves scenes in BEV space and demonstrates a significant capability to understand the global context

Ablation Study

Ablations on the proposed modules of BEV-TSR. In Table 2, we validate our proposed modules for model's performance on nuScenes-Retrieval. It is clear that incorporating each module facilitates the understanding of complex traffic scenarios and leads to performance gain in retrieving. Specifically, the BEV space learning (BEV) significantly improves the baseline of 35.5% R@1 scene retrieval, which reflects the BEV providing an intuitive representation of the world, which is most desirable for text-scene retrieval. Then the Shared Cross-modal Embedding (SCE) achieves remarkable improvements of 5.73% R@1 scene retrieval by aligning the BEV and text modalities within a unified embedding space. It's noted that when without the SCE module, we utilize an MLP layer for feature mapping. The Knowledge Graph Prompting (KGP) and Caption Generation (CG) modules also obtained consist boosts. Moreover, combining these modules leads to further improvements, indicating that they effectively contribute to retrieving complex autonomous driving scenes.

Ablations on the BEV Encoder. In the first block of Table 3, we investigate different BEV encoders for converting 2D features into 3D space, including BEVDet (Huang et al. 2021) and BEVformer (Li et al. 2022). Our approach achieves favorable retrieval results with both decoders, which demonstrates our generalization ability. While BEVFormer introduces sequential temporal modeling with

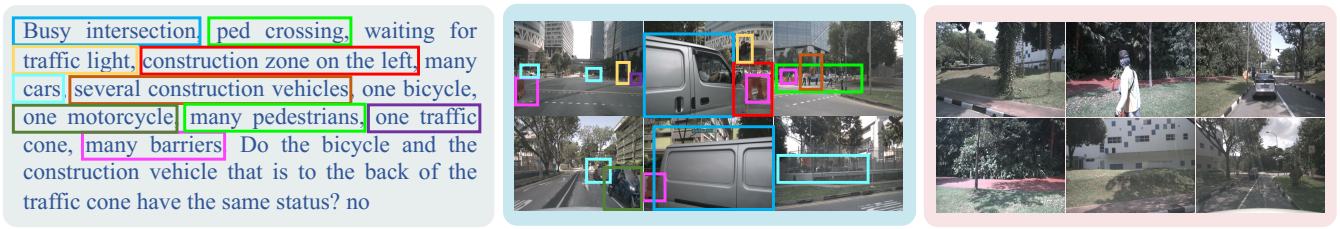

Figure 5: **R@1 scene retrieval results.** Our BEV-TSR achieves more precise retrieval results.

BEV	SCE	KGP	CG	Text Retrieval		Scene Retrieval	
				R@1	R@5	R@1	R@5
✗	✗	✗	✗	0.4846	0.9085	0.4644	0.9258
✓	✗	✗	✗	0.7875	0.9757	0.8194	0.9812
✓	✓	✗	✗	0.8352	0.9944	0.8431	0.9962
✓	✓	✓	✗	0.8469	0.9947	0.8557	0.9968
✓	✓	✓	✓	**0.8578**	**0.9954**	**0.8766**	**0.9971**

Table 2: Ablations on the proposed modules of BEV-TSR.

spatiotemporal transformers which efficiently extract scene information into BEV space, leading to optimal results.

Ablations on the Text Encoder. In the second block of Table 3, we conducted experiments to validate different text decoders for our proposed method, i.e., BERT (Devlin et al. 2018) and Llama2 (Touvron et al. 2023). Additionally, we also utilize LoRA (Hu et al. 2021) to fine-tune the Llama2 model. The results show that the Llama2 decoder is more effective than BERT in encoding relevant and accurate scene descriptions. Furthermore, fine-tuning the Llama2 models using the LoRA leads to gains of 11.64% in the R@1 scene retrieval. This improvement can be attributed to the fact that the pre-training task of Llama2 contains fewer traffic scenarios, fine-tuning tailors the models to the specific context of autonomous driving scenes.

Ablations on Knowledge Graph Prompting. In the third block of Table 3, we valid different knowledge graph embedding techniques, including TransE (Bordes et al. 2013), ConvE (Dettmers et al. 2018) and DistMult (Yang et al. 2014). By incorporating learned KGEs from all three models, we observed consistent improvements. This suggests that leveraging KGEs enhances the ability to understand and represent the relationships between entities and relations in the original KG, leading to improved retrieval performance. The results obtained with DistMult are slightly higher, leading us to select it as our default KGE extractor.

Ablations on Shared Cross-modal Embedding. In the last block of Table 3, we investigate the influence of the lantern shape of the shared cross-modal embeddings. The results reveal the larger lantern shape leads to better retrieval performance as they have the ability to adequately align two modalities' features. The shape of $k = 4096$ is sufficient to generate favorable retrieval results.

Module	Text Retrieval		Scene Retrieval	
	R@1	R@5	R@1	R@5
BEV Encoder				
BEVDet (Huang et al. 2021)	0.7955	0.9929	0.8021	0.9917
BEVFormer (Li et al. 2022)	**0.8578**	**0.9954**	**0.8766**	**0.9971**
Text Encoder				
BERT (Devlin et al. 2018)	0.6409	0.9129	0.5594	0.8915
Llama2 (Touvron et al. 2023)	0.7244	0.9472	0.7030	0.9507
w/ LoRA (Hu et al. 2021)	**0.7875**	**0.9757**	**0.8194**	**0.9812**
Knowledge Graph Prompting				
TransE (Bordes et al. 2013)	0.8487	0.9914	0.8559	0.9914
ConvE (Dettmers et al. 2018)	0.8501	0.9928	0.8602	0.9956
Distmult (Yang et al. 2014)	**0.8578**	**0.9954**	**0.8766**	**0.9971**
Shared Cross-modal Embedding				
k = 1024	0.8487	0.9912	0.8602	0.9971
k = 2048	0.8473	0.9937	0.8689	0.9968
k = 4096	**0.8578**	**0.9954**	**0.8766**	**0.9971**

Table 3: Ablations on each component of BEV-TSR.

Conclusion

In this paper, we propose the novel BEV-TSR framework for text-scene retrieval in autonomous driving, which retrieves scenes in BEV space and demonstrates a significant capability to understand the global context and retrieve complex traffic scenarios. For text sentence representation, we leverage an LLM and incorporate knowledge graph embeddings to comprehensively understand complex textual descriptions, offering a higher level of semantic richness in language embedding. To align the features, we propose Shared Cross-modal Embedding, which utilizes a set of shared learnable embeddings to bridge the gap between the BEV features and language embeddings in different feature spaces. We also leverage a caption generation task to further enhance the alignment. Moreover, we establish a multi-level retrieval dataset, nuScenes-Retrieval, based on the nuScenes dataset, on which our BEV-TSR achieves state-of-the-art performance with remarkable improvement, and extensive ablation experiments demonstrate the effectiveness of our proposed method.

Acknowledgments

This work was supported in part by National Science and Technology Major Project No.2024YFE0203100, National Natural Science Foundation of China (NSFC) under Grants No.62476293, National Science and Technology Ministry Youth Talent Funding No. 2022WRQB002, Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Nansha Key RD Program under Grant No.2022ZD014, and supported by General Embodied AI Center of Sun Yat-sen University and Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Guangzhou 510006, China.

References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Chen, L.; Sima, C.; Li, Y.; Zheng, Z.; Xu, J.; Geng, X.; Li, H.; He, C.; Shi, J.; Qiao, Y.; et al. 2022. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, 550–567. Springer.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Fu, D.; Lei, W.; Wen, L.; Cai, P.; Mao, S.; Dou, M.; Shi, B.; and Qiao, Y. 2024. LimSim++: A Closed-Loop Platform for Deploying Multimodal LLMs in Autonomous Driving. *arXiv preprint arXiv:2402.01246*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Li, H.; Sima, C.; Dai, J.; Wang, W.; Lu, L.; Wang, H.; Zeng, J.; Li, Z.; Yang, J.; Deng, H.; et al. 2023a. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, L.; Shao, W.; Dong, W.; Tian, Y.; Yang, K.; and Zhang, W. 2024. Data-Centric Evolution in Autonomous Driving: A Comprehensive Survey of Big Data System, Data Mining, and Closed-Loop Technologies. *arXiv preprint arXiv:2401.12888*.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2023c. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1486–1494.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023d. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35: 10421–10434.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lin, X.; Pei, Z.; Lin, T.; Huang, L.; and Su, Z. 2023. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, 531–548. Springer.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023b. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, 2774–2781. IEEE.

Long, A.; Yin, W.; Ajanthan, T.; Nguyen, V.; Purkait, P.; Garg, R.; Blair, A.; Shen, C.; and van den Hengel, A. 2022. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6959–6969.

Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; and Yan, J. 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1): 23–79.

Ma, Y.; Wang, T.; Bai, X.; Yang, H.; Hou, Y.; Wang, Y.; Qiao, Y.; Yang, R.; Manocha, D.; and Zhu, X. 2022. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*.

Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; and Han, B. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, 3456–3465.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K.; Tomizuka, M.; and Zhan, W. 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2023. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *arXiv preprint arXiv:2305.14836*.

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1655–1668.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Tong, W.; Sima, C.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; et al. 2023. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8406–8415.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3621–3631.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023b. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17850–17859.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wickramarachchi, R.; Henson, C.; and Sheth, A. 2020. An evaluation of knowledge graph embeddings for autonomous driving data: Experience and practice. *arXiv preprint arXiv:2003.00344*.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9433–9443.