Out-of-Label Hazard Detection for Autonomous Driving: Fusing Optical Flow, Depth, Proximity, and Scene Description

Weiqiang Zeng
Independent Researcher
*Corresponding author:zhwiqg953@gmail.com

***Abstract*—In this paper, we address the challenge of improving hazard detection in autonomous driving systems, particularly in scenarios where labeled data is scarce or unavailable. This issue is critical in real-world applications, where diverse and unpredictable driving situations make it difficult to label every potential hazard accurately. Recently, the Challenge of Out-of-Label (COOOL) benchmark has been introduced at WACV2025 to promote research on this challenge. To tackle this issue, we present a novel method that integrates a Bootstrapping Language-Image Pretraining (BLIP)-based scenario generation framework with a threshold-based hazard scoring system, thereby enhancing both scenario comprehension and detection accuracy within the benchmark. By incorporating robust driver state logic, bounding box analysis, and BLIP-generated scenario descriptions, our method initially achieves a 40% performance score. Building upon this foundation, we further integrate depth maps and optical flow to improve hazardous object discrimination, resulting in an additional 20% performance improvement. This culminates in a final score of 63% on the public benchmark leaderboard and 50% on the private leaderboard. To foster continued advancements in autonomous driving research, we will make all code and visualization tools publicly available.**

***Index Terms*—out-of-label, optical flow, depth maps, BLIP, image caption,hazard detection**

I. Introduction

With the rapid advancement of computer vision technologies [1]–[4], perception tasks in autonomous driving have evolved from fundamental 2D object detection [5]–[7], optical flow [8]–[10], and depth estimation [11]–[13] to more complex scene understanding through video anomaly detection. Recent breakthroughs in large-language Models (LLMs) [14]–[16]and Vision-Language Models(VLMs) [17]–[19] have demonstrated remarkable zero-shot reasoning capabilities, enabling LLMs to generate high-quality semantic interpretations without domain-specific training. These features give VLMs unique advantages in autonomous driving systems: effectively detecting road obstacles and identifying potential risk zones in driving scenarios through interpretable semantic descriptions. Such multi-modal (image to text) provides intuitive risk assessment references by establishing a bidirectional mapping between drive sense understanding and natural language generation, significantly enhancing decision-making transparency and reliability. Consequently, semi-supervised learning,few-shot learning, and zero-shot generative with multi-modal perception

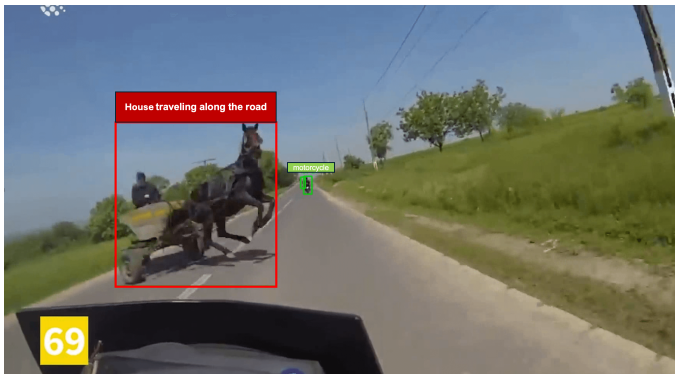

Fig. 1. A simplified result of our approach is displayed on the selected frame from one of the test videos. The colors represent the hazard state of each object: red indicates hazardous objects, and green indicates safe objects.

technologies have emerged as crucial research directions for improving driver-sense adaptability and safety redundancy in autonomous driving systems. While existing autonomous driving systems demonstrate remarkable proficiency in detecting predefined object categories (e.g., vehicles, pedestrians) within conventional benchmarks like KITTI, nuScenes and Waymo, their reliance on closed-set annotation paradigms creates critical safety blind spots. Current datasets predominantly focus on nominal driving scenarios, where over 98% of annotated objects fall within 20 common categories according to nuScenes statistics. According to NHTSA reports, this paradigm leaves systems fundamentally unprepared for Out-of-Distribution (OOD) hazards - unexpected objects and scenarios that account for 62% of real-world collision incidents. Such vulnerabilities manifest particularly in handling exotic biological entities (e.g., kangaroos crossing Australian highways), amorphous obstacles (e.g., wind-blown debris), and edge-case interactions (e.g., pedestrians emerging from visual occlusions), where traditional perception pipelines frequently fail to trigger appropriate emergency responses.

This study is based on the "Out-of-Label Hazards in Autonomous Driving (COOOL)" benchmark [20], a multi-modal dataset of high-resolution videos captured from real-world driving scenarios. COOOL is specifically designed to address the critical but underexplored challenge of detecting

out-of-distribution (OOD) hazards, which are categorized into three types: 1) Exotic biological threats (e.g., kangaroos, wild boars), 2) Unpredictable inanimate hazards (e.g., drifting plastic bags, smoke occlusion), and 3) Abnormal interactions with standard objects (e.g., erratic pedestrians). To deal with this problem, we propose the following methods:

- **Multi-modal Hazard Filtering:** Establish a priori conditions and optical flow and depth estimation to identify potential hazards based on motion discontinuity and spatial proximity.
- **Zero-Shot Categorization:** Use a CLIP-driven big model to classify filtered objects into predefined risk tiers without requiring task-specific training.
- **Causal Scene Interpretation:** Employ Vision Language Models (VLMs) to generate spatiotemporally grounded captions that explain the evolution of hazards (e.g., "A dog crossing the street").

II. Related Work

A. Optical Flow

Optical flow characterizes the perceived motion patterns between consecutive frames, representing the displacement vector field induced by relative motion between the observer and scene elements. This spatiotemporal signal provides critical cues for anticipating emerging threats in dynamic environments. Recent advancements in autonomous safety systems have increasingly leveraged optical flow for enhanced risk prediction and collision awareness. FlowNet 2.0 [21]established significant improvements in both estimation accuracy and computational efficiency, enabling real-time extraction of dense motion vectors. Building upon this [22] integrated optical flow with Occupancy Networks to predict the trajectories of dynamic obstacles, thus generating collision-free paths by incorporating vehicle kinematic constraints. In a similar vein, [23] developed a model that predicts Time to Collision (TTC) and optical flow from monocular images, identifying potential collision areas through feature clustering and motion analysis. Their model uses optical flow and TTC within a 65ms temporal window to assess collision risk. To further address challenges such as varying illumination, Wang et al. [24] fused monocular optical flow with stereo depth cues, successfully reducing optical flow errors by 50% compared to previous unsupervised methods.

B. Zero-Shot Image Classification

Recent advancements in vision-language pretraining have transformed open-vocabulary zero-shot learning. Pioneered by OpenAI's CLIP [25], which aligns 400 million image-text pairs into a unified embedding space through contrastive learning, this approach enables semantic transfer to unseen categories via natural language prompts. Building on this, ALIGN [26] further enhances multi-modal alignment by training on noisy web-scale data (1.8 billion pairs), demonstrating improved robustness in cross-modal retrieval tasks. In object detection, VILD [27] innovatively distills knowledge from CLIP-style classifiers into two-stage detectors like Mask R-CNN, effectively detecting rare categories using only base-class annotations. This highlights the possibility of open-vocabulary detection without relying on novel-class training data. Prompt engineering has also emerged as a key enabler for zero-shot adaptation. Methods like CoOp [28] optimize learnable context vectors to guide pre-trained vision language models (VLMs) toward downstream tasks, leading to a noticeable improvement in performance across multiple datasets. Further works like CoCoOp [18] introduced conditional prompt tuning, dynamically adjusting prompts based on image content, significantly reducing the domain gap on unseen classes.

C. Vision-Larger Language Models

The success of Vision Transformers (ViT) [29] and large-language Models (LLMs) has led to advances in cross-modal learning. ViT is used to extract hierarchical image features and then mapped into the textual embedding space of LLMs through alignment layers. For example, LLaVA [30]shows how aligning ViT outputs (D=1024) with LLM token dimensions (D=4096) using linear transformation enables visual question answering with minimal instruction tuning. Parameter-efficient fine-tuning [31] techniques have become essential for efficiently adapting models to new tasks. These include adapter-based tuning, which uses lightweight modules to adapt models with minimal parameter changes (e.g., VL-Adapter [32] tunes less than 1% of the total parameters), and Q-Former mechanisms, like those in BLIP [33], [34], where query vectors attend to key visual regions, speeding up convergence. These methods can deal with many challenges, including bridging the modality gap between ViT's grid-based features and LLM's sequential embeddings and ensuring efficient knowledge transfer by updating only the adapter parameters, making them suitable for tasks like autonomous hazard perception.

III. Method

As Fig 2,our approach begins by utilizing a priori knowledge to screen potential hazardous objects based on optical flow and depth information. These objects are then identified and categorized through zero-shot image captioning, allowing the model to recognize and classify hazards without requiring task-specific training. Finally, we use a vision language model to generate captions and categorize dangerous objects in each frame.

A. Multi-modal Hazard Filtering

We establish a prior assumption based on the intuition that larger and closer objects pose a greater danger. we design a hazard scoring mechanism defined as

$$score = \frac{bounding\ box\_size}{dist\_to\_center} \tag{1}$$

where objects with higher scores are considered more hazardous. This integrated scoring system enhances the accuracy of hazard assessment by prioritizing the highest-scoring object as the primary threat.We employ optical flow estimation for

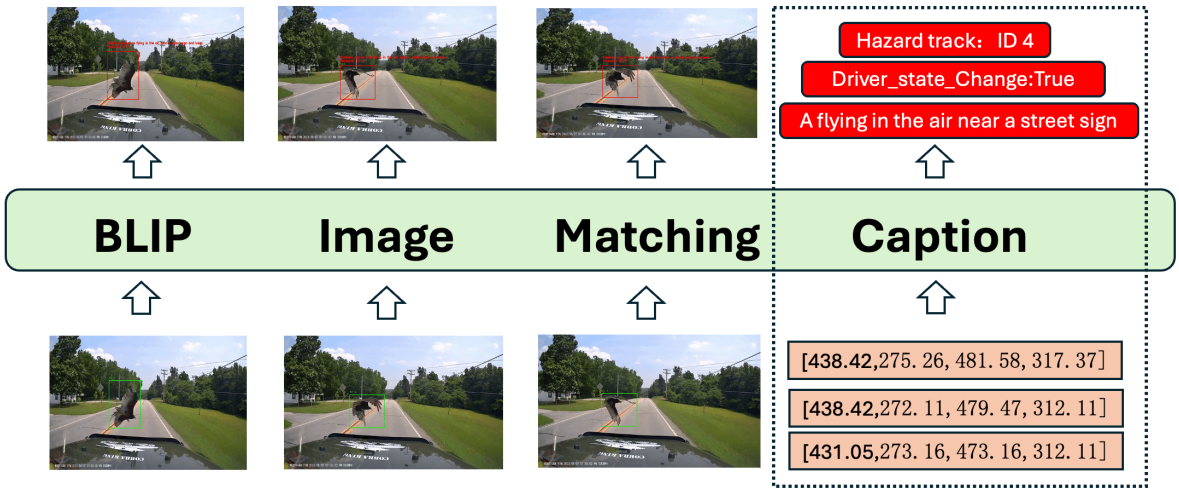

Fig. 2. Illustration of the proposed framework. BLIP, an advanced visual language model, is employed for image matching and captioning tasks to identify objects, determine potential hazards, and generate descriptions. Green boxes indicate bounding boxes with track IDs within the COOOL dataset.

TABLE I
COMPARISON OF PROCESSING TIMES FOR THE LINEAR REGRESSION AND THE SCORING MECHANISM IN DIFFERENT PROCESSING MODES ON THE COOOL DATASET.

Method	Processing Mode	Single Frame Time	Total Time
Linear	Single-threaded CPU	1 ms	4,320 s
	GPU Accelerated	0.01 ms	43.2 s
Scoring mechanism	Single-threaded CPU	0.01 ms	43.2 s
	GPU Accelerated	0.0001 ms	0.432 s

small objects and animals to capture how objects change instantaneously between consecutive frames. In dynamic environments, the optical flow field assists in identifying hazardous regions within a scene by scoring motion every five frames to assess whether the current driving state is potentially dangerous. Additionally, we incorporate monocular depth estimation in low-light conditions to predict scene depth. By analyzing variations in the depth map, we effectively distinguish moving objects and identify potential hazards, thereby enhancing the accuracy of hazard detection. The visualization of optical flow estimation and depth estimation is shown in Fig 4.

B. Zero-shot Image classification

For the identified hazardous objects, we extract them using the bounding boxes (bounding box) provided in the dataset and perform zero-shot image classification. However, relying solely on the bounding box may result in a loss of contextual information, making classification more challenging. To address this issue, we apply a 20% padding around the target image, ensuring that contextual cues are incorporated into the zero-shot model. For classification, we utilize OpenAI's CLIP ViT-B/16 [25] model and select the top 10 predicted categories with the highest probabilities as the final results.

C. Image Caption

We first employed a zero-shot classification method to process the input images, thereby identifying potentially hazardous objects in the scenes. Next, we used the BLIP model to generate detailed descriptions of the classified hazard objects. This model leverages the strengths of both visual information and large-language models to automatically image caption that accurately correspond to the characteristics of the hazardous objects. Meanwhile, by utilizing the frame-level label information provided in the dataset, we precisely located the keyframes containing the hazardous objects and conducted scene understanding on these frames. Based on the scene analysis results, we further examined the specific labels and attributes of the hazardous objects to formulate more accurate descriptions.

IV. DATASET

A. Annotation

The COOOL benchmark, entitled "Challenge Of Out-Of-Label" in Autonomous Driving, comprises 200 high-resolution dashcam videos that have been meticulously annotated by human labelers. The objective of this benchmark is to identify objects of interest and potential roadway hazards in Figure 1 . The range of potential hazards is extensive, including but not limited to exotic animals (e.g., birds, houses, dogs), unusual or unpredictable objects (e.g., plastic bags, smoke), and more common roadway threats (e.g., cars, pedestrians).

The annotation files illustrated in support object detection bounding boxes and follow the common object detection annotation format, providing us with $x_{\min}$, $x_{\max}$, $y_{\min}$, and $y_{\max}$ coordinates.

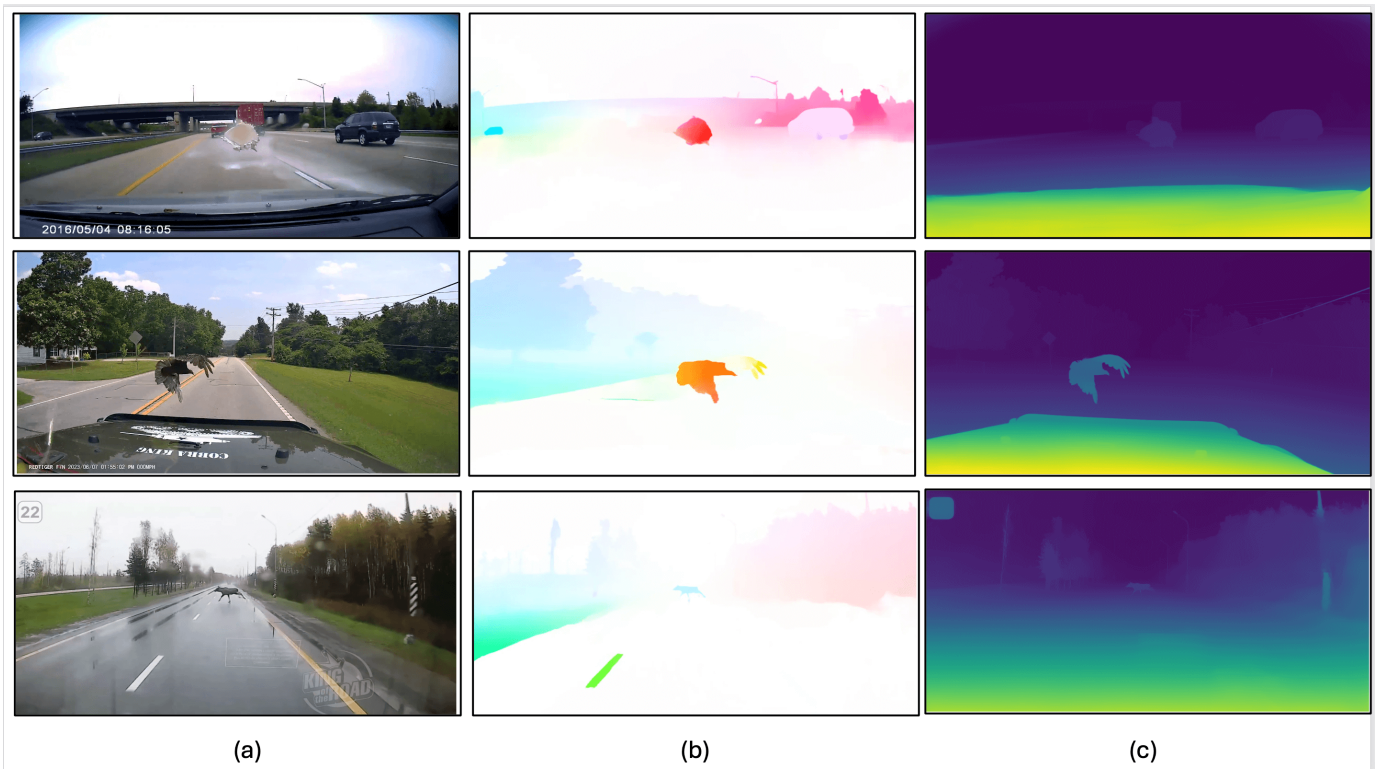

Fig. 3. The above images present the visualization of optical flow estimation and depth estimation. (a) is the original frame from the dataset, (b) is the optical flow estimation, and (c) is the depth map estimation.

TABLE II
CONSOLIDATED OBJECT DATA WITH OBJECT NAMES, ORDERED BY TRACK ID. ATTRIBUTES ARE INTENTIONALLY LEFT AS EMPTY BRACES ("{}") AT THIS STAGE. THIS TABLE MERGES CHALLENGE OBJECT DATA AND TRAFFIC SCENE DATA INTO ONE, WITH OBJECT NAMES ADDED.

Track ID	bounding box (Bounding Box)	Attributes	Object
0	[183.62, 497.99, 211.16, 538.2]	{}	traffic scene
1	[387.95, 457.78, 664.29, 686.97]	{}	challenge
2	[861.45, 576.45, 913.67, 648.1]	{}	challenge
3	[1047.92, 526.23, 1065.11, 542.62]	{}	traffic scene
4	[1050.36, 544.48, 1058.68, 567.64]	{}	traffic scene
5	[52.2, 656.7, 104.45, 700.1]	{}	challenge

B. Evaluation metrics

The COOOL competition evaluation metrics are intended to balance the three aspects of hazard detection. Datasets provide systems with a list of bounding boxes and the raw video, which enables diverse approaches to these challenges. In order to predict which potential hazards are genuinely hazardous, the accuracy of predictions is computed based on the maximum between the number of ground truth hazards and the number of predicted hazards.Let N_{gt} be the number of ground-truth hazards, N_{pred} be the number of predicted hazards, and N_{correct} be the number of correct hazard predictions. To penalize over-prediction, we use:

$$A_{\text{detection}} = \frac{2\,N_{\text{correct}}}{N_{\text{gt}} + N_{\text{pred}}}\,. \tag{2}$$

By adding the total number of hazards to the total number of guesses, algorithms that over-predict hazards are penalized, thus avoiding the inflation of accuracy through lucky guesses. For hazard descriptions, a similar approach is adopted, but here we only check whether the class label is included in the description, which is a binary evaluation. In Hazard Description Accuracy,For each hazard description, define the indicator function:

$$d_i = \begin{cases} 1, & \text{if hazard object will be explain,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

If there are N hazards to evaluate, then the description accuracy is:

$$A_{\text{description}} = \frac{1}{N}\sum_{i=1}^{N} d_i\,. \tag{4}$$

In the context of driver reactions, accuracy is determined based on the ground truth labels for each frame, thereby ascertaining whether the driver has reacted to the hazard. The overall evaluation metric is the macro-averaged accuracy of these three measures. For Driver Reaction Accuracy Let R_t be the ground-truth reaction label at frame t, and $\hat{R}_t$ be the predicted reaction label at frame t.Evaluated over T frames, the reaction accuracy is:

$$A_{\text{reaction}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\hat{R}_t = R_t\}\,, \tag{5}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function (1 if true, 0 otherwise).

Overall Evaluation,The overall metric is the macro-average of the three accuracies:

$$A_{\text{overall}} = \frac{1}{3}\Big(A_{\text{detection}} + A_{\text{description}} + A_{\text{reaction}}\Big). \quad (6)$$

V. Results and discussion

In the benchmark has not yet provided relevant label information, we use Kaggle's evaluation metrics as an indicator of our model's performance. As TABLE III showed that the gradual integration of various information modules significantly enhanced the overall performance. Initially, when only the CLIP model was employed, the system achieved an accuracy of merely 23%, indicating that relying solely on single-modal visual feature extraction is insufficient to capture the critical information of hazardous objects in complex driving scenes. By adopting the BLIP model, the accuracy slightly increased to 26%, demonstrating that BLIP possesses certain advantages in sense understanding and image captioning. However, it's still hard to capture the dynamic changes of the scene or analyze them in low-light conditions. Furthermore, when the BLIP model was combined with the Optical Flow estimation and scoring method, the accuracy improved to 42%, which validates the important role of incorporating motion information to capture dynamic changes between consecutive frames and enhance detection performance. Ultimately, our method further integrated depth map information to provide an in-depth depiction of the scene's geometric structure, elevating the reach to 63%. These results show the advantages of a multi-modal information fusion process in hazardous object detection.

TABLE III
Performance comparison of methods with component usage indicated by (✓) .

Method	CLIP	BLIP	Optical Flow	depth map	Score
Baseline	✓				23%
		✓			26%
		✓	✓		42%
Ours		✓	✓	✓	63%

Furthermore, the accuracy is further enhanced to 28% by incorporating a speed threshold, which improves predictions of driver state changes. By introducing a scoring strategy to evaluate the danger level of objects based on the inverse of their bounding box size and position relative to the center, the accuracy reaches 63%. These findings underscore the importance of integrating prior knowledge and adopting precise danger assessment methods to enhance prediction accuracy. A visualization of this approach is provided in Fig 4.

In addition, as shown in TABLE I, the threshold-based approach is 10 times faster than linear regression. This significant improvement enables the model to detect potential hazards and respond more quickly, which is a key factor in ensuring the real-time performance and safety of the autonomous driving system.

TABLE IV
COOOL challenge benchmark

#	Team name	$A^{public}_{reaction}$	$A^{private}_{reaction}$
1	Duong Anh Kiet	0.78453	0.57261
2	PiVa AI	0.68993	0.51772
3	Impish	0.63794	0.51596
4	**Ours**	0.63792	0.50599
5	Parisa Hatami	0.54599	0.48967
6	TeamCV	0.55705	0.44401
7	PMM_UTCU	0.43161	0.44020
8	Mahdi Abbariki	0.56956	0.37568
9	Nachiket Kamod	0.43368	0.31733
10	Peace.LU	0.34695	0.31639

VI. Conclusion and future work

This paper presents the approach we adopted in the COOOL Autonomous Driving Challenge, which requires the automatic detection of hazardous objects in driving scenarios without language annotations, as well as the generation of corresponding natural language descriptions. This task imposes stringent demands on existing vision-language models. To tackle this challenge, we propose a BLIP-based solution that integrates prior knowledge, optical flow, and depth estimation. Furthermore, we implement a fine-tuning strategy for large-language models by adjusting parameters such as vertex sampling, temperature, and competition degree. These improvements effectively enhance the overall performance of the model. Ultimately, our method significantly boosts accuracy, achieving a rate of 63%.As the TABLE IV Since the official paper for this competition has not yet been published, a direct comparison with other methods is not currently possible. However, our approach has demonstrated strong performance in experiments, indicating its competitive potential for this task.

In the future, we aim to explore advanced models such as LLaMA [35] and GPT-4.0 [15]. We plan to leverage chain-of-thought prompting to enhance the model's inference capabilities, enabling deeper semantic understanding and logical reasoning. Additionally, we intend to extend the model's capabilities to comprehend video data, allowing it to capture dynamic information and temporal relationships in driving scenarios. These advancements will further improve the model's performance and interpretability, contributing to the safe development of autonomous driving technology.

References

[1] C. Feng, B. Bačić, and W. Li, "Sca-lstm: A deep learning approach to golf swing analysis and performance enhancement," in *International Conference on Neural Information Processing*. Springer, 2025, pp. 72–86.

[2] B. Bačić, C. Feng, and W. Li, "Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation," *ISBS Proceedings Archive*, vol. 42, no. 1, p. 60, 2024.

[3] J. Wang, S. Wang, and Y. Zhang, "Deep learning on medical image analysis," *CAAI Transactions on Intelligence Technology*, vol. 10, no. 1, pp. 1–35, 2025.

[4] Y. Zhong and S. H. Lee, "Gazesymcat: A symmetric cross-attention transformer for robust gaze estimation under extreme head poses and gaze variations," *Journal of Computational Design and Engineering*, vol. 12, no. 3, pp. 115–129, 2025.

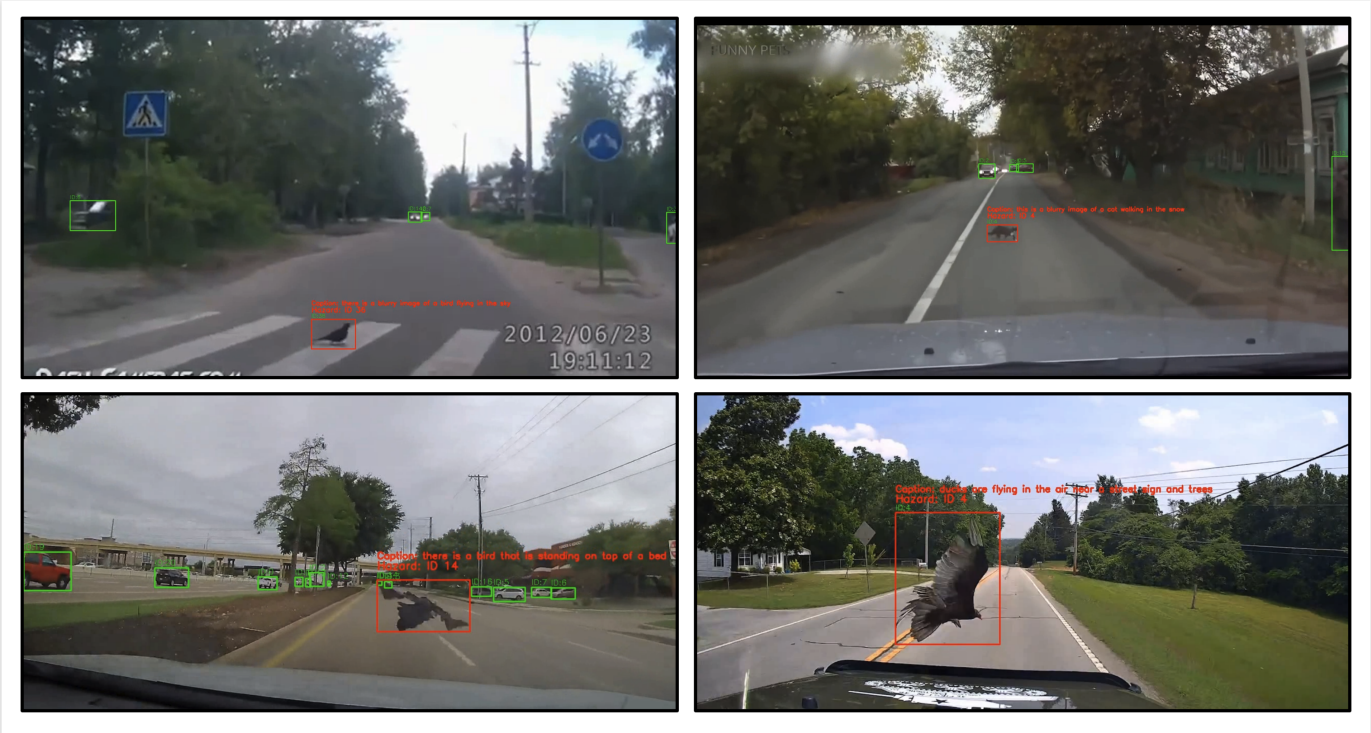

Fig. 4. Sample predictions from our model in the dataset. Green boxes indicate bounding boxes for detected objects, while red boxes highlight hazardous targets within the scene.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[7] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 965–16 974.

[8] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[9] D. Sun, D. Vlasic, C. Herrmann, V. Jampani, M. Krainin, H. Chang, R. Zabih, W. T. Freeman, and C. Liu, "Autoflow: Learning a better training set for optical flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 093–10 102.

[10] S. Khairi, E. Meunier, R. Fraisse, and P. Bouthemy, "Efficient local correlation volume for unsupervised optical flow estimation on small moving objects in large satellite images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 440–448.

[11] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in neural information processing systems*, vol. 18, 2005.

[12] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[13] Z. Li, X. Wang, X. Liu, and J. Jiang, "Binsformer: Revisiting adaptive bins for monocular depth estimation," *IEEE Transactions on Image Processing*, 2024.

[14] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of chatgpt on benchmark datasets," 2023. [Online]. Available: https://arxiv.org/abs/2305.18486

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, B. He, S. Jiang, and B. Dong, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2303.16421

[17] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5579–5588.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[19] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[20] A. K. AlShami, A. Kalita, R. Rabinowitz, K. Lam, R. Bezbarua, T. Boult, and J. Kalita, "Coool: Challenge of out-of-label a novel benchmark for autonomous driving," *arXiv preprint arXiv:2412.05462*, 2024.

[21] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[22] R. Mahjourian, J. Kim, Y. Chai, M. Tan, B. Sapp, and D. Anguelov, "Occupancy flow fields for motion forecasting in autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5639–5646, 2022.

[23] C. Li, Y. Qian, C. Sun, W. Yan, C. Wang, and M. Yang, "Ttc4mcp: Monocular collision prediction based on self-supervised ttc estimation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 244–250.

[24] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching

videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8071–8081.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[26] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021. [Online]. Available: https://arxiv.org/abs/2102.05918

[27] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[28] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[31] A. P. Gema, P. Minervini, L. Daines, T. Hope, and B. Alex, "Parameter-efficient fine-tuning of llama for the clinical domain," *arXiv preprint arXiv:2307.03042*, 2023.

[32] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5227–5237.

[33] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[34] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[35] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.