Proceedings of The 2023 IEEE International
Conference on Real-time Computing and Robotics
July 17-20, 2023, Datong, China

Knowledge Distillation on Driving Intention Generator: Learn Human-like Semantic Reasoning

Hejun Wang, Jiajun Dong, Yunkai Wang, Yanmei Jiao, Rong Xiong†

***Abstract*— Semantic information plays a significant role in autonomous vehicles. However, it has remained open how to economically compute and efficiently incorporate semantic information into other frequently applied processes, such as construction of local navigation map, driving decision making and trajectories planning. To address this issue, we introduce Knowledge Distillation to convey semantic reasoning capability. Following the existing intention-guided framework, a semantically incorporated module associates given navigation instruction with acquired visual perception to synthesize driving intention. Our major contribution is to teach the driving intention generator how to propagate considering semantics implicitly. Specifically, we construct a semantic supervisor identical to the driving intention generator. The supervisor transfers semantic reasoning ability to the generator during training process, with which the generator is expected to perform stably and reliably even in extremely sophisticated surroundings or under frequently inaccurate navigation instruction. Through exhaustive experimental validation, our model demonstrates a superior robustness and adaptability compared to the state-of-the-art. Finally, we explain the intrinsic mechanism of our model with visualization.**

I. Introduction

Autonomous driving technology is a highly researched topic recently. Current studies concentrate on effective, efficient and economic integration of navigation, drivers' intention and temporal sensing information before making driving decision. However, the majority of state-of-the-art autopilots lack human-like semantic reasoning process, for it considerably consumes computational resource and the effective integration of semantics with other perceptions remains an open issue.

Researchers [1]–[5] have raised various networks which are widely considered to be end-to-end, vision-based and resistant to slight disturbance, for it is proven that the traditional modular paradigm combining perception, localization and navigation relies on the high precision of each module and results in a limited applicability. P. Cai et al. [1] proposes an end-to-end architecture that directly extracts features from vision then generates trajectory. Given that features extracted from vision lack interpretability and the network fails to visibly refine navigation instructions that contain errors, the reliability and stability of their proposed method remain controversial. H. Ma et al. [6], [7] pioneered a framework to solve this problem. In their framework, a generative module is employed to synthesize driving intention associating high-hierarchy navigation instructions with acquired first-person perspective vision images. Synthetic driving intention is expected to maintain parallel alignment with lanes and provide appropriate orientation even when instructions come with errors. This work improves interpretability and could overcome the instability of navigation instructions and gives relatively accurate local instructions downstream. However, under extreme situations such as sudden errors occurred in navigation instructions or intricately coupled traffic scenarios, the intention generator is prone to making mistakes that may lead the vehicle off-course or towards the roadside.

Incorporating human-like semantic recognition is considered as a viable solution. Works in [8]–[11] attempted to assimilate semantics into identification of buildings, lanes, pedestrians, etc., as well as recognizing relationships between different objects, resulting in better local map construction and driving trajectory planning that are significantly safer and more reliable. While, the achievement of high accuracy in semantic reasoning comes at the cost of increased computational requirements. This trade-off leads to the development of overly complex and inefficient intelligent systems.

In this paper, we propose to augment an autonomous driving system capability to conduct semantic reasoning without incurring excessive computational overhead or sophisticated integration operations. To accomplish this objective, we introduce *Knowledge Distillation* [12] into the driving intention generator. This is facilitated by incorporating a semantic supervisor, which is trained with privileged information, and serves to transfer semantic reasoning inclinations to other modules.

Specifically, we adopt the framework of H. Ma et al. [6], but enhance the generative module a capability to conduct human-like semantic reasoning. Taking inspiration from the work of W. Chen et al. [13]–[15], we construct a semantic supervisor that imparts semantic knowledge during the training process. The semantic supervisor is structurally identical to the driving intention generator while it receives privileged information in the form of well semantically segmented images, in contrast to the raw images fed to the driving intention generator. We apply the aforementioned approaches in CARLA simulation [16] to ascertain the effectiveness. The trained model demonstrates remarkable robustness to poor navigation and adaptability across various applications, achieving high levels of performance. Additionally, we visualize the sensitive areas of the model, similar to Grad-CAM

Hejun Wang, Jiajun Dong, Yunkai Wang, Rong Xiong are with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China.

Yanmei Jiao is with the School of Information Science and Engineering, Hangzhou Normal University, Hangzhou, China.

† Corresponding author, rxiong@zju.edu.cn

979-8-3503-2718-2/23/$31.00 ©2023 IEEE

approach proposed by Selvaraju et al. [17]. This provides compelling evidence that our proposed strategy effectively enhances the target module with human-like semantic reasoning capability.

To summarize, the main contributions of this paper include the following:

- We propose a novel strategy based on knowledge distillation to complement the driving intention generation module with human-like semantic reasoning capabilities.
- Validations on CARLA demonstrate that our proposed method can provide trained models with unprecedented robustness and adaptability.
- We visualize the effect of semantic distillation, providing a strong explanation for the mechanism of our deep neural networks.

II. Related Works

A. End-to-End Methods

After the paradigm has emerged, there are more and more researchers engaged in developments of end-to-end methods [18]. In contrast to earlier ideas of mapping directly from a wide variety of perceptions to actions [5], [18], current models prefer optimizing for interpretability, portability, reliability and so on [1], [2], [7], [19].

For instance, VTGNet [1] produces discrete points of trajectory instead of a sequence of actions. Separation of trajectory generation and vehicle control is a wise choice for greater interpretability, higher safety and easier training. What's more, HDNet [19] is designed to excel in association differing data from varying sensors in order to draw a high-definition map for surroundings. Chen et al. [20] center their work on the treatment of perception and achieve very promising results.

Of great relevance to our work, H. Ma et al. [6] devise a deterministic *conditional Generative Adversarial Networks* (cGAN) [21], that given low-precise navigation instruction and real-time image, a local driving intention could be computed by a UNet [22]. Our study endeavors to replicate the tactics pioneered by them while joining knowledge distillation techniques with the purpose of amplifying the semantic acumen of the intention generator.

B. Semantic Annotation

Indeed, semantic annotation is among the focal points [23]–[26] where vision cameras demonstrate their superiority over conventional sensors. Certain studies [27]–[30] focus on improving the robustness of traffic semantic recognition in complex environmental conditions. Others [31]–[33] aim to confirm security and reliability.

Involving semantic annotation in vehicle navigation has been extensively researched. In 2016, a deep deconvolutional network has been proposed by P. Alcantarilla et al. [34]. This network is designed for semantic pixel-wise change detection in images of urban scenes to update navigation maps. V. Murali et al. [35] utilize semantic landmarks for precise navigation and succeeded in accuracy on GPS-denied navigation solutions. S. Yasmin et al. [36] avoid severe incidents with semantic segmentation to detect small obstacles.

What's more, effective integration of semantics with other perception remains a formidable challenge. As a breakthrough, Hu et al. [37] define a semantic-based generic representations and apply Graph Neural Networks to mimic human-brain reasoning. It is inspired that innovatively designing a model with human-like semantic reasoning ability is superior to simply configuring an isolated semantic recogniser.

C. Model Distillation

In 2016, Hinton et al. [12] have demonstrated a knowledge distillation method that could enhance the performance of nearly any machine learning algorithm. This approach achieves this without incurring excessive computational costs or being overly complex to implement. Furthermore, A. Romero et al. [38] claim that not only the final output of the teacher model, but also the intermediate outputs of its hidden layers can be utilized to supervise the student model. This practice is even more effective since these hidden layers usually have fewer parameters yet contain deeper and more essential features.

This technique has been proven efficient in the field of autonomous driving technology [14], [15], [38]–[40]. In this study, we symmetrically design a teacher model, called semantic supervisor, and train it with the student model simultaneously.

III. Methodology

Following the existing intention-guided framework [6], we utilize a well-trained generative module to integrate high-hierarchy navigation instructions and real-time perceptions to synthesize driving intentions to meet the demands of downstream trajectory planner. Then to enhance the generative module a capability to conduct human-like semantic reasoning, we introduce knowledge distillation into the learning stage by adding a semantic supervisor. Specifically, the semantic supervisor is constructed identically to the generator. Through end-to-end learning, the generator can indirectly learn semantic reasoning mechanism from the semantic supervisor. In addition, a binary discriminator $\mathcal{D}$ is designed specifically to support the deterministic cGAN [6]. The overarching system architecture is depicted in Fig. 1.

A. Modeling

The intention generator $\mathcal{G}$ assumes a central role by translating a raw images I and a screenshot of a navigation instruction$N$ into driving intentions $R^{\mathcal{G}}$:

$$R^{\mathcal{G}} = \mathcal{G}(I, N), \tag{1}$$

The semantic supervisor $\mathcal{S}$ is structurally equivalent to $\mathcal{G}$, while interfaces with semantically annotated images E, yielding driving intentions $R^{\mathcal{S}}$:

$$R^{\mathcal{S}} = \mathcal{S}(E, N) \tag{2}$$

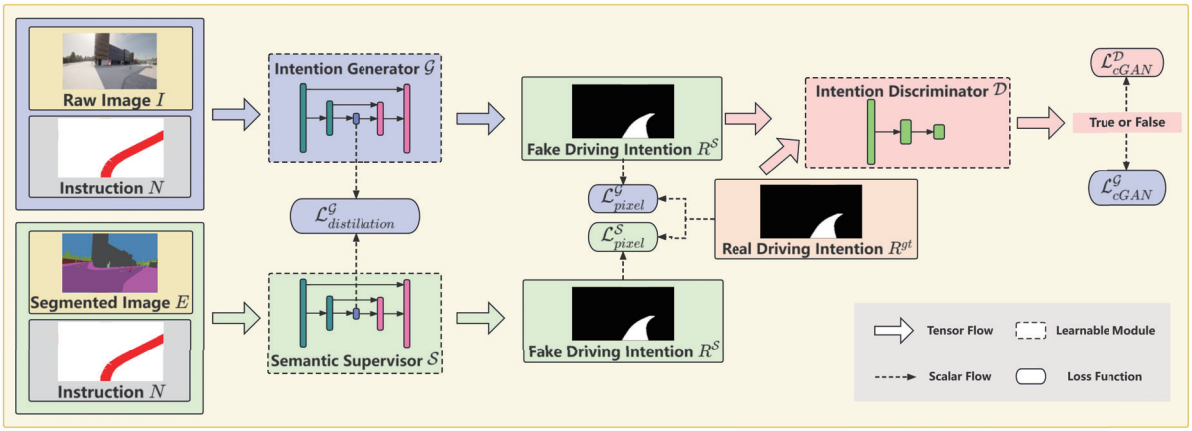

Fig. 1. Our proposed method comprises three learnable modules: an intention generator $\mathcal{G}$ (in purple), a semantic supervisor $\mathcal{S}$ (in green) and an intention discriminator $\mathcal{D}$ (in pink). The intention generator, employing a UNet structure, takes a raw vision and a screenshot of navigation instruction as input, and is trained to synthesize a local driving intention represented as a grayscale image. Simultaneously, the semantic supervisor, constructed identically to the intention generator, takes a well semantically segmented version of the input to produce the same grayscale image as the intention generator. To train these models, an L1-distance loss function is employed to compare the pixel-wise similarity of their outputs with corresponding ground truth, which are drawn from authentic trajectories. From the idea of model distillation, a loss of middle layers in the UNet for the intention generator is added. This additional loss term promotes the learning of semantic annotation. Moreover, the intention discriminator is trained to differentiate between authentic intentions and those generated ones. The adversarial loss is also employed to improve the overall quality of the model. In summary, the loss function for the intention generator comes from three primary terms: comparison against the ground truth, core differences, and adversarial learning. The method of semantic distillation is incorporated to effectively conveying knowledge from a privileged model (the semantic supervisor) to the intention generator.

Both $R^{\mathcal{G}}$ and $R^{\mathcal{S}}$ are expected to approximate the ground truth $R^{\mathbf{gt}}$:

$$\mathcal{L}^{\mathcal{G}}_{pixel} = \mathbb{E}_{I,N,R}[||R^{\mathbf{gt}} - \mathcal{G}(I,N)||_1] \tag{3}$$

$$\mathcal{L}^{\mathcal{S}}_{pixel} = \mathbb{E}_{E,N,R}[||R^{\mathbf{gt}} - \mathcal{S}(E,N)||_1] \tag{4}$$

Here the L1 distance is employed to guarantee steep gradients around the zero point [41].

B. Adversarial Training

As a conditional binary classifier [21], the intention discriminator $\mathcal{D}$ outputs a boolean variable that serves as an arbiter to evaluate the reality and feasibility of synthetic driving intention. During model training, we optimize $\mathcal{D}$ with a cross-entropy loss given by:

$$\begin{aligned}\mathcal{L}^{\mathcal{D}}_{cGAN} = &\mathbb{E}_{I,N,R}[log(1 - \mathcal{D}(I,N,R))] \\ &+ \mathbb{E}_{I,N}[log(1 + \mathcal{D}(I,N,\mathcal{G}(I,N)))]\end{aligned} \tag{5}$$

Both corresponding and non-corresponding navigation instruction screenshots, N and $N^{'}$ are given to strengthen resistance to incorrect instruction of the intention generator. The intention generator, in turn, aims to optimize reality and feasibility of its outputs via minimizing the loss function given by:

$$\begin{aligned}\mathcal{L}^{\mathcal{G}}_{cGAN} = &\mathbb{E}_{I,N}[log(1 - \mathcal{D}(I,N,\mathcal{G}(I,N)))] \\ &+ \mathbb{E}_{I,N'}[log(1 - \mathcal{D}(I,N^{'},\mathcal{G}(I,N^{'})))]\end{aligned} \tag{6}$$

C. Semantic Distillation

To train a model capable of mimicking human thought processes, we develop a mechanism known as semantic distillation. Both the intention generator $\mathcal{G}$ and the semantic supervisor $\mathcal{S}$ adopt UNet, following H. Ma et al. [6]'s method. In UNet, there is a low-dimensional tensor located at the middle layer called "core" C. Theoretically, the core contains the most significant and critical information required for accurate generation. In our methodology, we exploit the core as a focal point for supervision.

The only difference between $\mathcal{S}$ and $\mathcal{G}$ is that $\mathcal{S}$ takes semantically annotated images E as inputs, rather than raw images I. Explicitly, the $\mathcal{S}$ is the very one which conduct semantic reasoning. However, a semantically annotated image could be hardly obtained easily, meaning that $\mathcal{S}$ is useless on practical deployment. While one of the innovations lies here: we deliberately construct a mechanism that $\mathcal{S}$ will teach $\mathcal{G}$ how to propagate. In other words, the intention generator $\mathcal{G}$ will learn a capability to conduct semantic reasoning from $\mathcal{S}$ during training process.

Doing so required minimizing the difference between the core tensors $C^{\mathcal{G}}$ and $C^{\mathcal{S}}$, thus resulting in $\mathcal{G}$ learning to think like $\mathcal{S}$ in intermediate process.

Here comes a loss function for $\mathcal{G}$:

$$\mathcal{L}^{\mathcal{G}}_{distillation} = \mathbb{E}_{I,N,R}[||C^{\mathcal{G}} - C^{\mathcal{S}}||_2] \tag{7}$$

Above all, we set a group of loss functions for the intention generator $\mathcal{G}$, the semantic supervisor $\mathcal{S}$ and the intention discriminator $\mathcal{D}$:

$$\mathcal{L}^{\mathcal{G}} = \mathcal{L}^{\mathcal{G}}_{cGAN} + \lambda_1 \mathcal{L}^{\mathcal{G}}_{pixel} + \lambda_2 \mathcal{L}^{\mathcal{G}}_{distillation} \tag{8}$$

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}^{\mathcal{S}}_{pixel} \tag{9}$$

$$\mathcal{L}^{\mathcal{D}} = \mathcal{L}^{\mathcal{D}}_{cGAN} \tag{10}$$

where λ_1 and λ_2 are manually set parameters. The third term in (8) incentivizes the intention generator $\mathcal{G}$ to extract

and incorporate semantic details from raw images, thereby advancing its proficiency in semantic reasoning.

IV. Experiment

To prove semantic distillation affects, we conduct a series of experiments to examine robustness against disturbance on instruction, competency to correct instruction and adaptability to various driving condition. Visualization is also utilized for mechanism analysis.

A. Experimental Setup

Collecting Datasets: Train sets, validation sets and test sets are all collected on CARLA, an open source autonomous driving simulator. We drive a virtual car equipped with a high-resolution camera, semantic camera and low-precise GPS to collect total about 20 sets of traffic scenes at 10 Hz. Positional information is also logged to synthesize the ground truth R^{gt} of the synthetic intentions $R^{\mathcal{G}}, R^{\mathcal{S}}$.

Comparative Models: For the sake of better persuasion, we select three baseline methods to show comparative results with the proposed method. These models are following:

- *cGAN*: H. Ma et al. [6] utilize cGAN for adversarial training. We train this model without semantic distillation. The others are applied equivalently.
- *UNet*: UNet is a frequently applied module proven fast and effective [22]. We train this model without semantic distillation and adversarial training. The others are applied equivalently.
- *CNN*: VTGNet [1] extracts visual features based on a sequence of bottleneck convolutional layers, so that we design an intention generator based on convolutional networks. We train this model without semantic distillation while the others are applied equivalently.
- *Ours*: This is an intention generator trained in our proposed semantic distillation methods.

B. Comparative Results

Disturbance on Instruction: To simulate the instability of navigation instruction in real-world applications and investigate how those intention generators react to this noise, we set up three levels of challenges: *Easy*, *Moderate* and *Hard*, with the probability of disturbances in navigation instruction ranging from 0.20 to 0.90.

We employed three main metrics for evaluation:

- IoU, Intersection over Union. It is defined as the intersection area divided by the union area, of the generated intention and the real intention.
- CLC, Central Line Coverage. It is defined as the proportion of the length of central line of synthetic intention in the real intention area.
- α, the absolute difference of angle between straight line joining the beginning and end of the central line of the synthetic intention and that of the ground truth.

The results are presented in Table I. The experimental results show that the proposed method performs similarly to the comparison method in the *Easy* case, while it performs best and has a clear advantage in the presence of large noise such as *Moderate* and *Hard*. These results show that the proposed method is more robust against disturbances, which are often prevalent and non-negligible in practice.

Instruction Corrections: Fig. 2 illustrates the behaviors of different models under several unreliable navigation instructions. It is obvious that *cGAN*, *UNet*, and *CNN* perform poorly in instruction correction, as they directly synthesize driving intentions without taking practical situations into consideration. In addition, they completely fail when the instruction vanishes. In contrast, *Ours*, thanks to the semantic distillation employed, is able to synthesize driving intention with integration of visual semantics, and correction of incorrect instructions is therefore achieved.

Multi-weather Behaviors: Fig. 3 displays the behaviors of the different models in various condition, including midnight, fog, and rainstorm. It is apparent that *Ours* exhibits strong adaptability in various conditions, compared to the others whose results may shatter, swell, etc., leading to mistakes in the subsequent process. This demonstrates that our model is not only less dependent on instruction, but can also be adapted to other noisy application scenarios.

In conclusion, our proposed semantic distillation approach has been proven to be effective in enhancing the modules with capability to conduct semantic reasoning. The results reveal that our model outperforms other models in most tested scenarios, showing higher robustness against poor navigation conditions and adaptability to various applications.

C. Mechanism Analysis

To reveal the intrinsic mechanism of the proposed method in which we utilize a visualization method that we derived the average output as derivatives of the input variables to obtain heat maps of the absolute gradients, overlaid on the corresponding original inputs, as shown in Fig. 4. These results implies that *cGAN* and *CNN* are insensitive to specific

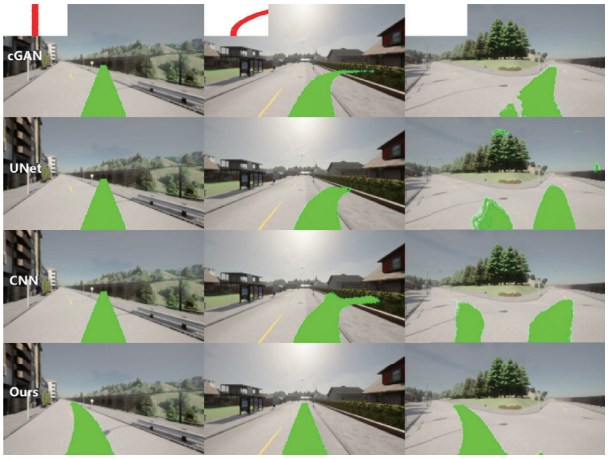

Fig. 2. The display shows the behaviors of *cGAN*, *UNet*, *CNN* and *Ours* under unreliable navigation, with the navigation map presented in the top left corner of the first row. It is evident that except for *Ours*, these models lack the ability to correct navigation. These findings provide compelling evidence for our model's ability to perform human-like semantic reasoning, as evidenced by its active avoidance of obstacles and strict driving intention to lane-keeping.

TABLE I

RESULTS OF DISTURBANCE ON INSTRUCTION

level		**Easy**			**Moderate**			**Hard**	
model	IoU(%)	CLC(%)	α(deg)	IoU(%)	CLC(%)	α(deg)	IoU(%)	CLC(%)	α(deg)
cGAN [6]	72.84	**92.01**	**5.47**	65.63	82.76	11.15	54.48	70.05	20.04
UNet [22]	78.14	91.56	11.46	70.92	83.97	14.04	54.94	68.24	20.54
CNN [1]	**79.90**	91.40	9.20	68.98	79.75	15.13	53.24	64.70	24.30
Ours	76.38	89.08	9.30	**76.66**	**89.40**	**8.19**	**77.97**	**90.23**	**7.40**

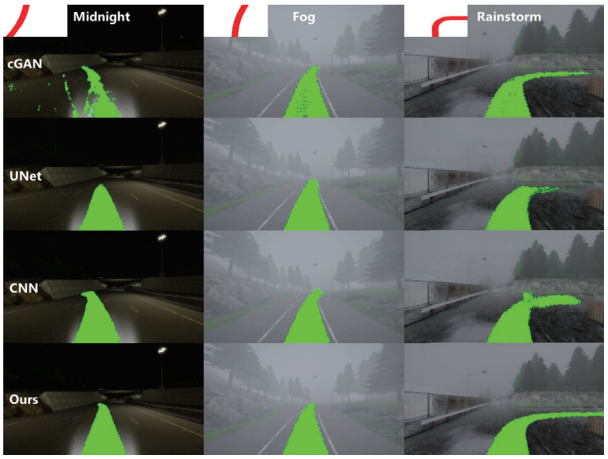

Fig. 3. The models' performances under different weather conditions are displayed, with a navigation map located in the top left corner of the first row. The performance of each model can be clearly observed in corresponding weather conditions. It is worth noting that models based on the previous methods often produce errors such as breaking, swelling, and appearing hollow, while ours demonstrates stronger adaptability in these tested conditions.

visual features, and *UNet* seems sensitive to the edges of the ideal output that are neither barriers nor lane edges. Moreover, they are more sensitive to instruction than vision. In contrast, our semantically learning model is activated by crucial details, e.g. contours of roads, buildings on the opposite side and other vehicles. This is the result of semantic distillation and the reason why our model performs better.

Nevertheless, our approach may also attend to irrelevant areas, leading to an unsteady output and potentially undermining model performance. Thus, it warrants further research to optimize the activation mechanism and minimize the impact of non-relevant information.

V. CONCLUSION

In this paper, we propose a semantic distillation mechanism that enables the generator to indirectly learn semantic reasoning capabilities under the supervision of a semantic supervisor. Our experiments reveal that the proposed method is more robust and adaptive to unreliable and unstable navigation instruction, as well as varying weather conditions. We also visualize the mechanism underlying the semantic reasoning in heat map, providing guidelines for subsequent network design. Future research is also required to confirm feasibility on real vehicles and to reduce the sensitivity of latent visual noise.

REFERENCES

[1] P. Cai, Y. Sun, H. Wang, and M. Liu, "Vtgnet: A vision-based trajectory generation network for autonomous vehicles in urban environments," 2020.

[2] A. Filos, P. Tigas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" 2020.

[3] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 731–13 737.

[4] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 853–17 862.

[5] P. Cai, S. Wang, Y. Sun, and M. Liu, "Probabilistic end-to-end vehicle navigation in complex dynamic environments with multimodal sensor fusion," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4218–4224, 2020.

[6] H. Ma, Y. Wang, L. Tang, S. Kodagoda, and R. Xiong, "Towards navigation without precise localization: Weakly supervised learning of goal-directed navigation cost map," 2019.

[7] Y. Wang, D. Zhang, J. Wang, Z. Chen, Y. Li, Y. Wang, and R. Xiong, "Imitation learning of hierarchical driving model: From continuous intention to continuous trajectory," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2477–2484, 2021.

[8] M. Colley, B. Eder, J. O. Rixen, and E. Rukzio, "Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445351

[9] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for the development of automated vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1813–1820.

[10] R. Juric and O. Madland, "Semantic framework for creating an instance of the ioe in urban transport: A study of traffic management with driverless vehicles," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1–8.

[11] D. K. Dewangan and S. P. Sahu, "Road detection using semantic segmentation-based convolutional neural network for intelligent vehicle system," in *Data Engineering and Communication Technology*, K. A. Reddy, B. R. Devi, B. George, and K. S. Raju, Eds. Singapore: Springer Singapore, 2021, pp. 629–637.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[13] W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, and Z. Wang, "Fasterseg: Searching for faster real-time semantic segmentation," 2020.

[14] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," 2019.

[15] S.-I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," 2019.

[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, oct 2019. [Online]. Available: https://doi.org/10.1007%2Fs11263-019-01228-7

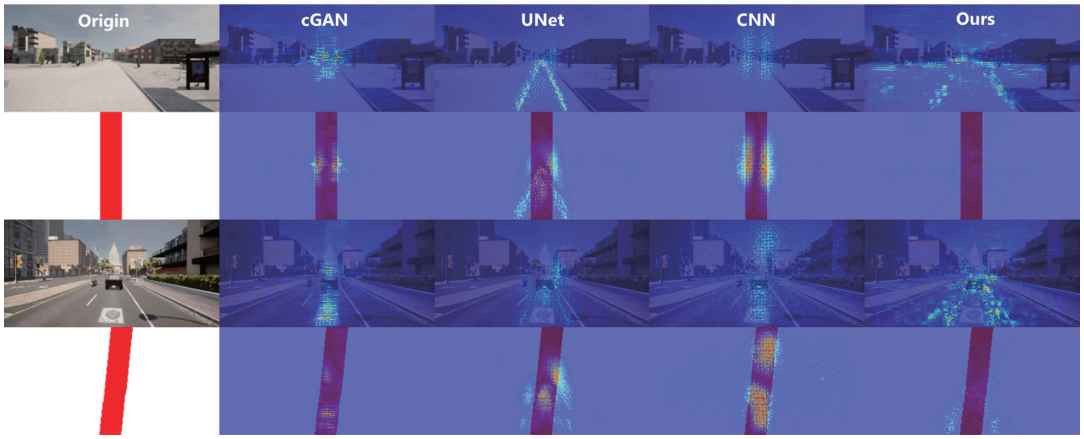

Fig. 4. The heat maps depict the absolute gradient of the average output with respect to the inputs, and they are overlaid on the original inputs. Warmer patches indicate areas that play a more decisive role in the output, while colder areas suggest the model is not sensitive to those regions. Compared to others, whose heat maps seems unattached to anything specific in the vision, *Ours* is captivated with driveway lines, road junctions, barriers on the opposite side of the road, which explains why *Ours* has capabilities to capture semantic features and correct wrong instruction. Nevertheless, this also suggests that our approach may be susceptible to interference from vision. More research is needed to clarify the transmission routes and mitigate the effects of disturbances. Despite this, our approach still holds great potential for enhancing the interpretability and performance of deep learning models in various fields.

[18] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," 2016.

[19] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," 2022.

[20] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2722–2730.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[23] H. Ma, R. Xiong, Y. Wang, S. Kodagoda, and L. Shi, "Towards open-set semantic labeling in 3d point clouds : Analysis on the unknown class," *Neurocomputing*, vol. 275, pp. 1282–1294, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231217315904

[24] H. Ma, L. Shi, S. Kodagoda, and R. Xiong, "A semantic labeling strategy to reject unknown objects in large scale 3d point clouds," in *2016 35th Chinese Control Conference (CCC)*, 2016, pp. 7070–7075.

[25] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," 2019.

[26] A. H. Khan, M. Munir, L. van Elst, and A. Dengel, "F2dnet: Fast focal detection network for pedestrian detection," 2022.

[27] Z. Pan, T. Emaru, A. Ravankar, and Y. Kobayashi, "Applying semantic segmentation to autonomous cars in the snowy environment," 2020.

[28] T. Liu and T. Stathaki, "Faster r-cnn for robust pedestrian detection using semantic segmentation network," *Frontiers in Neurorobotics*, vol. 12, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2018.00064

[29] K. Yang, W. Zhang, C. Li, and X. Wang, "Accurate location in dynamic traffic environment using semantic information and probabilistic data association," *Sensors*, vol. 22, no. 13, p. 5042, Jul 2022. [Online]. Available: http://dx.doi.org/10.3390/s22135042

[30] M. Yan, J. Wang, J. Li, K. Zhang, and Z. Yang, "Traffic scene semantic segmentation using self-attention mechanism and bi-directional gru to correlate context," *Neurocomputing*, vol. 386, pp. 293–304, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219317072

[31] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[32] V. Besnier, D. Picard, and A. Briot, "Learning uncertainty for safety-oriented semantic segmentation in autonomous driving," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 3353–3357.

[33] N. A. Surobhi, Y. Ma, and A. Jamalipour, "A semantic agglomerative traffic management framework for ubiquitous public safety networks," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, 2011, pp. 26–30.

[34] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Autonomous Robots*, vol. 42, pp. 1301–1322, 2016.

[35] V. Murali, H.-P. Chiu, S. Samarasekera, and R. T. Kumar, "Utilizing semantic visual landmarks for precise vehicle navigation," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–8.

[36] S. Yasmin, M. Y. Durrani, S. Gillani, M. Bukhari, M. Maqsood, and M. Zghaibeh, "Small obstacles detection on roads scenes using semantic segmentation for the safe navigation of autonomous vehicles," *Journal of Electronic Imaging*, vol. 31, no. 6, p. 061806, 2022. [Online]. Available: https://doi.org/10.1117/1.JEI.31.6.061806

[37] Y. Hu, W. Zhan, and M. Tomizuka, "Scenario-transferable semantic graph reasoning for interaction-aware probabilistic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 23 212–23 230, 2022.

[38] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2015.

[39] M. Phuong and C. H. Lampert, "Towards understanding knowledge distillation," 2021.

[40] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," 2016.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.