Benchmarking Formal Verification for Autonomous Driving in the Wild

Yonggang Luo [1] Jinyan Ma [1] Sanchu Han [1] Lecheng Xie [1]

Abstract

The verification of the security of neural networks is cruicial, especially for the field of autonomous driving. Although there are currently benchmarks for the verification of the robustness of neural networks, there are hardly any benchmarks related to the field of autonomous driving, especially those related to object detection and semantic segmentation. Thus, a notable gap exists in formally verifying the robustness of semantic semantic segmentation and object detection tasks under complex, real-world conditions. To address this, we present an innovative approach to benchamark formal verification for autonomous driving perception tasks. Firstly, we propose robust verification benchmarks for semantic segmentation and object detection, supplementing existing methods. Secondly, and more significantly, we introduce a novel patch-level disturbance approach for object detection, providing a more realistic representation of real-world scenarios. By augmenting the current verification benchmarks with our novel proposals, our work contributes towards developing a more comprehensive, practical, and realistic benchmarking methodology for perception tasks in autonomous driving, thereby propelling the field towards improved safety and reliability. Our dataset and code used in this work are publicly available [1] [2].

1. Introduction

The advent of autonomous driving technologies has ushered in a new paradigm for transportation, offering promise for improved safety and efficiency. Yet, they also pose significant challenges, particularly in perception tasks which are fundamental to their safe operation. For example, by accurately segmenting the drivable road surface from other areas, the vehicle can know where it can go. In the other case, the autonomous vehicle needs to detect pedestrians to ensure that it stops or slows down in time to avoid collisions. Reliable perception necessitates precise object detection and semantic segmentation under varied and often unpredictable real-world conditions (Shen et al., 2022), where semantic segmentation is the task of clustering parts of images together which belong to the same object class (Thoma, 2016) and object detection is the task of identifying objects in the image along with their localizations and classifications (Chahal & Dey, 2018).

Formal verification, grounded in rigorous mathematical and logical principles, has been identified as a potent mechanism for assuring the safety and performance of neural network. Many formal approaches are already able to verify variants of classification tasks (Anderson et al., 2019; Botoeva et al., 2020; Dathathri et al., 2020; Fazlyab et al., 2022; Katz et al., 2019; Mohapatra et al., 2020; Ruan et al., 2018; Singh et al., 2019; Tjeng et al., 2017; Tran et al., 2020; Zhang et al., 2018; Xu et al., 2020; Salman et al., 2019; Xu et al., 2021; Wang et al., 2021; Zhang et al., 2022b;a; Ferrari et al., 2022; Henriksen & Lomuscio, 2020; 2021; Henriksen et al., 2021; Khedr et al., 2020; Bak, 2021; Brix & Noll, 2020). The verification of safety and robustness specification of neural network controlled systems is explored by many works (Huang et al., 2019; Ivanov et al., 2019; Tran et al., 2019). Furthermore, the performance of image-based controllers is discussed by concatenating the generator network with the control network (Katz et al., 2021). However, only few work focuses on the formal approach for verifying semantic segmentation and object detection networks robustness using reachability analysis (Tran et al., 2021). Moreover, existing formal verification methods often rely on assumptions of ideal operational environments, creating a potential divergance from the often unpredictable conditions encountered in real-world scenarios.

[1] AI Lab, Chongqing Changan Automobile Ltd., Chongqing, China. Correspondence to: Yonggang Luo <luoyg3@changan.com.cn>.

Presented at the 2^{nd} Workshop on Formal Verification of Machine Learning, co-located with the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA., 2023. Copyright 2023 by the author(s).

[1] https://github.com/pomodoromjy/vnn-comp-2022-Carvana-unet

[2] https://github.com/pomodoromjy/vnncomp-2023-CCTSDB-YOLO

Given this context, although we have many benchmarks for formal verification methods (Bak et al., 2021; Müller et al., 2023; Brix et al., 2023), there is still a significant and unexplored need to benchmark formal verification methods for autonomous driving system in the wild. This approach allows for a more realistic assessment of perception neural networks' robustness under challenging real-world conditions, while also facilitating calibration of verification tools to better mirror reality. The motivation behind our work arises from the necessity to enhance the verification benchmarks for object detection and semantic segmentation tasks and to better align them with actual autonomous driving scenarios.

In this paper,we present a comprehensive approach to benchmark formal verification for autonomous driving perception tasks. Our primary contributions are two-fold: firstly, we propose robust verification benchmarks for object detection and semantic segmentation tasks. Secondly, and more importantly, we introduce a patch-level disturbance approach for object detection tasks, mirroring the complexities of real-world scenarios in a more realistic manner. Although adversarial samples can effectively attack our percepttion models, in the real world, we seldom encounter disturbance patterns that exactly match adversarial samples. That is, it's nearly impossible to replicate pixel-level disturbances in the real world, so it's questionable whether we will encounter adversarial samples' interference patterns in the real world. On the contrary, patch-level disturbance patterns are a more common type of interference, and they are easier to replicate in the real world and are more likely to occur. For example, we only need to simply cut some black paper pieces to replicate the disturbance patterns we want to appear in the real world. By augmenting the existing verification benchmarks and proposing a novel patch-level disturbance approach, this work aims to provide a more comprehensive and practical benchmarking methodology for autonomous driving perception tasks, thereby advancing the field towards greater safety and reliability.

The paper is organized as follows: In Section 2, we provide a detailed relevent work. Section 3 introduces our semantic segmentation benchmark. In Section 4, we delve into our patch-level object detection benchmark. Finally, Section 5 presents the experiment of two benchmarks, and Section 6 concludes the paper with future work directions.

2. Related Work

Perception tasks in autonomous driving: Perception tasks hold a key role by playing a critical function in recognizing and understanding the various elements in the surrounding environments (Yurtsever et al., 2019). This understanding allows these perception tasks to extract vital semantic information necessary for safe and efficient driving. Such information includes the identification and detection of different road onjects. These could be pedestrians crossing the street, other vehicles in transit, or even potential obstacles that could hinder the smooth progress of the autonomous vehicle.

Moreover, object tracking is another crucial perception task, ensuring a continous understanding of the movement and position of surrounding entites. Another aspect of perception tasks involves semantic segmentation, a process that categorizes each pixel in an image to a particular class to help the vehicle better understand its environment. This not only includes road and off-road classification but also recognizes different lanes and traffic lights, aiding the vehicle's decision-making progress in different traffic scenarios.

These perception tasks rely heavily on the integration of multiple sensor inputs. These sensors typically include cameras, Light Detection and Ranging (LiDAR) systems and Radio Detection and Ranging (RADAR) sensors.The confluence of data from these diverse sensor systems feeds into the perception tasks, aiding the autonomous vehicle in understanding and navigating its surroundings efficiently and safely.

In this paper, we mainly focus on semantic segmentation and object detection tasks. Considering that the research in formal verification of neural networks is still unable to handle complex neural network models, we have simplified the model in our benchmarks, which is not intended for commercial mass production or practical use. At the same time, we are only considering the data of a single target in a single camera as the models' input. We will treat more complex perception models and multi-sensor inputs as future research directions.

Benchmarks: Well-known benchmarks for perception tasks are typically KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019) and Waymo (Sun et al., 2020) in autonomous driving area, however they were designed for general purpose instead of performing robustness evaluation. In order to conduct evaluations of robustness, recent studies have been actively either developing new benchmarks based on the existing autonomous driving datasets (Dong et al., 2023), or constructing new datasets that consist of road anomalies, or those that represent extrem weather conditions (Chan et al., 2021; Hendrycks et al., 2022; Li et al., 2022; Pinggera et al., 2016; Bijelic et al., 2020; Diaz-Ruiz et al., 2022; Pitropov et al., 2020).

In the field of robustness verification of neural networks, benchmarks are typically image classification tasks, though some recent studies have been actively proposing new benchmarks in many other tasks (Bak et al., 2021; Müller et al., 2023; Brix et al., 2023). However, to the best of our knowledge, so far there is no paper focusing on benchmarks which evaluate formal verification tools for perception tasks of autonomous driving in the wild.

In this paper, we propose two benchmarks which are related to autonomous driving scenarios, and we present a detailed description of the background to highlight our benchmarks' relevance and the characteristics of the verification problems. The ONNX format and the VNN-LIB format were adpoted for our benchmarks.

3. Semantic Segmentation Benchmark – Carvana Unet

The motivation behind our proposed benchmarks is primarily the predominant focus of existing networks in the literature on image classification. We perceive the need for more emphasis on aspects such as object detection or semantic segmentation, particularly in real-world scenarios such as autonomous driving. In this section, we introduce a new suite of simplified Unet (Ronneberger et al., 2015) benchmarks designed specifically for neural network verification on the Carvana dataset (Brian Shaler, 2017). To respond to the practicality of current verification tools and the intricate nature of semantic segmentation, we construct this new series of simplified Unet benchmarks (model one consists of four Conv2d layers followed by BatchNorm (BN) and ReLu; model two builds upon model one, adding an AveragePool layer and a Transposed Conv Upsampling layer). We believe that it's vital for tools to address more pragmatic architectures and consider this simplified Unet as a step in that direction.

Furthermore, the Carvana dataset, composed of 5088 images representing 318 cars (16 images per car), has been divided into a test set of 318 images (one per car) and a training set of the remaining 4700 images. The input images should be normalized to a [0, 1] range. Ground truth masks, generated by running the model on original images, assign either a 0 or 1 to each pixel. Our proposal is to select 16 images randomly for verification from those whose over 98.8 percent and 99.0 percent of pixels are predicted correctly by model one and model two respectively. The input size is [1, 4, 31, 47], where '1' corresponds to the batch size, '4' to the number of channels, '31' and '47' to the height and width of samples respectively. The first three channels signify RGB values of images, and the last channel denotes the model-produced mask used for computing the quantity of accurately predicted pixels by the model. The model output is the count of pixels predicted correctly by the model, juxtaposed with the model-produced mask.

4. Patch-level Object Detection Benchmark – CCTSDB YOLO

While the Carvana Unet benchmark in section 3 allows the application of neural network verification tools in autonomous driving scenarios, the pixel-level perturbation is still challenging to reflect the real-world situation. In this section, we are stepping up the challenge by introducing a new set of benchmarks for object detection within autonomous driving scenes. Given the practicality of current verification tools, we have modified Yolo-FastestV2 (Ma, 2021), based on a well-known end-to-end object detection framework Yolo. This architecture comprise backbone, neck, and head components.

To further alleviate computational burden, we have simplified the backbone and neck. For the head, we aim to facilitate single object detection while bypassing the need to conduct non-maximum suppression (NMS) operation within the model. To this end, we have replaced the box regression method with landmark regression for coordinate detection.

To the best of our knowledge, previous benchmarks were designed to test the model's digital world robustness. However, with an eye towards real-world practicality, we now suggest testing the model's robustness within the physical world. Specifically, we will supply an image with its corresponding label, as well as a fixed-size patch (either 1×1 or 3×3). Our goal is for the community to verify the model's robustness after applying the patch to any position within the image, all within the allocated time of specific time.

We utilized the training set from CCTSDB 2021 (Zhang et al., 2022c), which encompasses a total of 16356 images (26838 instances). Further division of all instances in a 9:1 ratio resulted in a training set comprising 23856 instances and a test set featuring 2982 instances. The input images and target coordinates need normalization within the range of 0 - 1. Targets are divided into three categories, signified by 0 (mandatory), 1 (prohibitory), and 2 (warning). We picked images with an intersection over union (IoU) greater than 0.5 and correct category classification from the test set. Eventually, 16 images will be selected at random for verification.

The model input consists of an array of 12296 elements, which include images (12288 elements), position (2 elements), and targets (6 elements). The model's single output is a combination of IoU between the predicted and actual bounding box, and the consistency of the predicted category with the actual category, as Equation (1).

$$output = IoU \times \begin{cases} 1, & pred\_cls = gt\_cls \\ 0, & pred\_cls \neq gt\_cls \end{cases} \tag{1}$$

If the final output for the input with the added patch is less

Table 1. Object detection accuracy for the dataset with/without patches.

MODEL	WITHOUT PATCH	WITH PATCH
MODEL1 (PATCH SIZE 1×1)	0.968	0.805
MODEL2 (PATCH SIZE 3×3)	0.978	0.253

than 0.5, the model is deemed non-robust for that patch. And vice versa.

5. Experiments

For the Carvana Unet benchmark, three formal verification tools (α, β Crown (Zhang et al., 2018; Xu et al., 2020; Salman et al., 2019; Xu et al., 2021; Wang et al., 2021; Zhang et al., 2022b;a), MN-BAB (Ferrari et al., 2022), and VeriNet (Henriksen & Lomuscio, 2020; 2021; Henriksen et al., 2021)) have been successfully applied to our benchmark in VNN-COMP 2022 (Müller et al., 2023). The number of instances that were solved by the different formal verification tools within a certain runtime for our benchmark is as illustrated in Figure 1. We expect more formal verification tools could be applied to the Carvana Unet benchmark in the future.

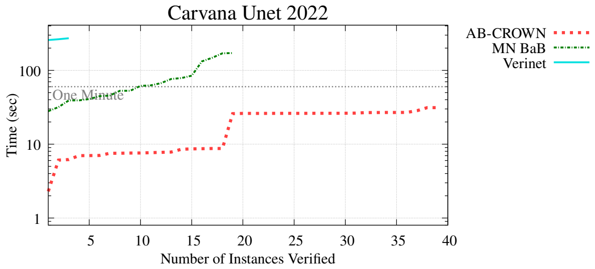

Figure 1. The number of instances that were solved by the different formal verification tools within a certain runtime for Carvana Unet in VNN-COMP 2022 (Müller et al., 2023).

For the CCTSDB YOLO benchmark, we further conduct extensive experiments to prove the following properties; (1) The simplified object detection model can still accurately identify targets within the dataset. (2) By adding patches randomly, we can ensure an anomalous detection in some images while maintaining correct detection in others, thus preventing situations where all data are either hold or violated. The object detection model and the dataset are same as the description in section 4. We summarize our experiment results in the Table 1.

As illustrated in the Table 1, the result shows that without the patch added, the model achieves successful detection rates of 0.968 and 0.978, indicating that its performance on our dataset has not significantly declined due to simplification. Moreover, after adding the patch, the detection rates drop to 0.805 and 0.253, ensuring that some data fails detection, thereby validating the effectiveness of our benchmark in evaluating the performance of formal verification tools (within a certain time limitation).

6. Conclusion

In this paper, we show how perception tasks' performance can be further connected with robustness verification field by benchmarking formal verification for autonomous driving in the wild. Specifically, we propose two benchmarks consist of the pixel-level semantic segmentation benchmark (Carvana Unet) and the patch-level object detection benchmark (CCTSDB YOLO). Experiments results demonstrate the effectiveness of the proposed benchmark for evaluating formal verification tools in autonomous driving perception tasks. Future work will take into account more real-world autonomous driving tasks (e.g., 3D object detection, object tracking and LiDAR localization) and more variants of attached patches.

References

Anderson, G., Pailoor, S., Dillig, I., and Chaudhuri, S. Optimization and abstraction: A synergistic approach for analyzing neural network robustness. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pp. 731–744, New York, NY, USA, 2019. Association for Computing Machinery.

Bak, S. nnenum: Verification of relu neural networks with optimized abstraction refinement. In *NASA Formal Methods Symposium*, pp. 19–36. Springer, 2021.

Bak, S., Liu, C., and Johnson, T. T. The second international verification of neural networks competition (vnn-comp 2021): Summary and results. *ArXiv*, abs/2109.00498, 2021.

Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., and Heide, F. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11679–11689, 2020.

Botoeva, E., Kouvaros, P., Kronqvist, J., Lomuscio, A., and Misener, R. Efficient verification of relu-based neural networks via dependency analysis. In *AAAI Conference on Artificial Intelligence*, 2020.

Brian Shaler, DanGill, M. M. M. P. W. C. Carvana image masking challenge, 2017.

Brix, C. and Noll, T. Debona: Decoupled boundary network analysis for tighter bounds and faster adversarial robustness proofs. *ArXiv*, abs/2006.09040, 2020.

Brix, C., Muller, M. N., Bak, S., Johnson, T. T., and Liu, C. First three years of the international verification of neural networks competition (vnn-comp). *ArXiv*, abs/2301.05815, 2023.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2019.

Chahal, K. S. and Dey, K. A survey of modern object detection literature using deep learning. *ArXiv*, abs/1808.07256, 2018.

Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R. Y., Salzmann, M., Fua, P., and Rottmann, M. Segmentmeifyoucan: A benchmark for anomaly segmentation. *ArXiv*, abs/2104.14812, 2021.

Dathathri, S., Dvijotham, K., Kurakin, A., Raghunathan, A., Uesato, J., Bunel, R. R., Shankar, S., Steinhardt, J., Goodfellow, I., Liang, P. S., and Kohli, P. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5318–5331. Curran Associates, Inc., 2020.

Diaz-Ruiz, C. A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., Chao, W.-L., Hariharan, B., Weinberger, K. Q., and Campbell, M. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21351–21360, 2022.

Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., and Zhu, J. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. *ArXiv*, abs/2303.11040, 2023.

Fazlyab, M., Morari, M., and Pappas, G. J. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, 67(1):1–15, 2022.

Ferrari, C., Mueller, M. N., Jovanović, N., and Vechev, M. Complete verification via multi-neuron relaxation guided branch-and-bound. In *International Conference on Learning Representations*, 2022.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. X. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022.

Henriksen, P. and Lomuscio, A. Efficient neural network verification via adaptive refinement and adversarial search. 2020.

Henriksen, P. and Lomuscio, A. Deepsplit: An efficient splitting method for neural network verification via indirect effect analysis. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2549–2555. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

Henriksen, P., Hammernik, K., Rueckert, D., and Lomuscio, A. Bias field robustness verification of large neural image classifiers. In *British Machine Vision Conference*, 2021.

Huang, C., Fan, J., Li, W., Chen, X., and Zhu, Q. Reachnn: Reachability analysis of neural-network controlled systems. *ACM Trans. Embed. Comput. Syst.*, 18(5s), oct 2019.

Ivanov, R., Weimer, J., Alur, R., Pappas, G. J., and Lee, I. Verisig: Verifying safety properties of hybrid systems with neural network controllers. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '19, pp. 169–178, New York, NY, USA, 2019. Association for Computing Machinery.

Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., Dill, D. L., Kochenderfer, M. J., and Barrett, C. The marabou framework for verification and analysis of deep neural networks. In Dillig, I. and Tasiran, S. (eds.), *Computer Aided Verification*, pp. 443–452, Cham, 2019. Springer International Publishing.

Katz, S. M., Corso, A., Strong, C. A., and Kochenderfer, M. J. Verification of image-based neural network controllers using generative models. *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–10, 2021.

Khedr, H., Ferlez, J., and Shoukry, Y. Effective formal verification of neural networks using the geometry of linear regions. *ArXiv*, abs/2006.10864, 2020.

Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.-Y., Liang, X., Li, Z., and Xu, H. Coda: A real-world road corner case dataset for object detection in autonomous driving. *ArXiv*, abs/2203.07724, 2022.

Liu, X., Yang, H., Liu, Z., Song, L., Chen, Y., and Li, H. H. Dpatch: An adversarial patch attack on object detectors. *arXiv: Computer Vision and Pattern Recognition*, 2018.

Ma, X. dog-qiuqiu/yolo-fastestv2: V0.2, August 2021.

Mohapatra, J., Weng, T.-W., Chen, P.-Y., Liu, S., and Daniel, L. Towards verifying robustness of neural networks against a family of semantic perturbations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 241–249, 2020.

Müller, M. N., Brix, C., Bak, S., Liu, C., and Johnson, T. T. The third international verification of neural networks competition (vnn-comp 2022): Summary and results, 2023.

Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., and Mester, R. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1099–1106, 2016.

Pitropov, M. A., Garcia, D. E., Rebello, J., Smart, M. H. W., Wang, C., Czarnecki, K., and Waslander, S. L. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40:681 – 690, 2020.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

Ruan, W., Wu, M., Sun, Y., Huang, X., Kroening, D., and Kwiatkowska, M. Global robustness evaluation of deep neural networks with provable guarantees for L0 norm. *CoRR*, abs/1804.05805, 2018.

Salman, H., Yang, G., Zhang, H., Hsieh, C.-J., and Zhang, P. A convex relaxation barrier to tight robustness verification of neural networks. *Advances in Neural Information Processing Systems*, 32:9835–9846, 2019.

Shen, J., Wang, N., Wan, Z., Luo, Y., Sato, T., Hu, Z., Zhang, X., Guo, S., Zhong, Z., Li, K., Zhao, Z., Qiao, C., and Chen, Q. A. Sok: On the semantic ai security in autonomous driving. *ArXiv*, abs/2203.05314, 2022.

Singh, G., Gehr, T., Püschel, M., and Vechev, M. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), jan 2019.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020.

Thoma, M. A survey of semantic segmentation. *ArXiv*, abs/1602.06541, 2016.

Tjeng, V., Xiao, K. Y., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2017.

Tran, H.-D., Cai, F., Diego, M. L., Musau, P., Johnson, T. T., and Koutsoukos, X. Safety verification of cyber-physical systems with reinforcement learning control. *ACM Trans. Embed. Comput. Syst.*, 18(5s), oct 2019.

Tran, H.-D., Bak, S., Xiang, W., and Johnson, T. T. Verification of deep convolutional neural networks using imagestars. In Lahiri, S. K. and Wang, C. (eds.), *Computer Aided Verification*, pp. 18–42, Cham, 2020. Springer International Publishing.

Tran, H.-D., Pal, N., Musau, P., Lopez, D. M., Hamilton, N., Yang, X., Bak, S., and Johnson, T. T. Robustness verification of semantic segmentation neural networks using relaxed reachability. In Silva, A. and Leino, K. R. M. (eds.), *Computer Aided Verification*, pp. 263–286, Cham, 2021. Springer International Publishing.

Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.-J., and Kolter, J. Z. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. *Advances in Neural Information Processing Systems*, 34, 2021.

Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.-W., Huang, M., Kailkhura, B., Lin, X., and Hsieh, C.-J. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33, 2020.

Xu, K., Zhang, H., Wang, S., Wang, Y., Jana, S., Lin, X., and Hsieh, C.-J. Fast and Complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2019.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4944–4953, Red Hook, NY, USA, 2018. Curran Associates Inc.

Zhang, H., Wang, S., Xu, K., Li, L., Li, B., Jana, S., Hsieh, C.-J., and Kolter, J. Z. General cutting planes for bound-propagation-based neural network verification. *Advances in Neural Information Processing Systems*, 2022a.

Zhang, H., Wang, S., Xu, K., Wang, Y., Jana, S., Hsieh, C.-J., and Kolter, Z. A branch and bound framework for stronger adversarial attacks of ReLU networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 26591–26604, 2022b.

Zhang, J., Zou, X., Kuang, L., Wang, J., Sherratt, R., and Yu, X. Cctsdb 2021: A more comprehensive traffic sign detection benchmark. *Human-centric Computing and Information Sciences*, 12:23, 05 2022c.

A. Network Details

In our benchmarks, we used a total of three networks, namely Unet_simp, Unet_upsample, and Yolo. Among them, Unet_simp and Unet_upsample correspond to benchmark Carvana Unet, while Yolo corresponds to benchmark CCTSDB YOLO. We have summarized the amount of parameters and the size of the models corresponding to these thress networks in the Table 2. The networks in benchmark Carvana Unet used operation such as Conv, BN, ReLu, AvgPool, ConvTranspose, etc., whereas the networks in benchmark CCTSDB YOLO used operations like Conv, BN, ReLu, MaxPool, interpolate, etc.

Table 2. Summary of the amount of parameters and model sizes of the three networks

NETWORKS	THE AMOUNT OF PARAMETERS	MODEL SIZES (M)
UNET_SIMP	149826	0.608
UNET_UPSAMPLE	330370	1.333
YOLO	144583	0.668

B. Implementation Details

For the benchmark Carvana Unet, we used the RMSprop optimizer, where the weight decay was set to 1×10^{-8} and the momentum was set to 0.9. We initialized the learning rate to 1×10^{-5}, with a decay strategy of ReduceLROnPlateau, where the mode was chosen as max and patience was set to 2. We trained it for a total of 5 epochs.

For the benchmark CCTSDB YOLO, we used the SGD optimizer, where the weight decay was set to 0.0005 and the momentum was set to 0.949. We initialized the learning rate to 0.001, with a decay strategy of MultiStepLR, where the milestones was set to an array as [150, 250] and gamma was set to 0.1. We trained it for a total of 300 epochs.