We are pleased to submit our revised manuscript entitled “Multi-Scale Grid Attention and Probabilistic Refinement for Accurate RoI-based Monocular 3D Object Detection” for consideration as an article in *IEEE Transactions on Intelligent Transportation Systems*. The original version of this manuscript was previously submitted to TITS and subsequently rejected(T-ITS-24-06-2169). However, we have thoroughly addressed the reviewers' comments and made significant modifications as follows:

1. The abstract and conclusion have been revised for greater clarity and specificity.
2. A future work section has been added to outline potential research directions that could not be included in this paper.
3. A latency comparison with various methods has been incorporated to demonstrate the real-time capability of our approach.
4. Additional ablation studies have been included to justify the use of the Multi-Scale Grid Attention (MSGA) mechanism over other attention methods.
5. All equations have been corrected to consistently end with appropriate punctuation (periods or commas).
6. We have updated the methods used for comparison to include more recent approaches from 2022 to 2024.
7. The title has been changed to better highlight the main contributions of the paper.
8. Long sentences in the manuscript have been split to improve readability.

However, there are a few reviewer comments that we did not address:

1. We did not include additional datasets because state-of-the-art (SOTA) monocular 3D object detection methods typically utilize three datasets: KITTI, nuScenes, and Waymo. While our experiments cover KITTI and Waymo, the nuScenes dataset, which utilizes six cameras, is more suited to multi-view methods, which fall beyond the scope of this paper. This has been explained in the experimental section.
2. We did not delve into denoising methods, as they are typically part of depth completion studies, which are outside the scope of our research.
3. We did not compare our method with point cloud-based approaches, as we believe these represent fundamentally different methodologies that are not directly comparable.
4. Visualization and explanations were already included in the original manuscript, but may have been overlooked by the reviewers. For example, Table VIII provides an ablation study on hyper-parameters, which appears to have been missed by Reviewer 3 in their 11th comment.

Abstract of the paper:

Monocular 3D object detection has gained increasing attention due to its cost-effectiveness and simplified setup. This study focuses on Region of Interest (RoI)-based monocular detectors. Previous approaches have treated all parts of the RoI equally, overlooking the fact that different

regions within the RoI hold varying levels of importance. Additionally, accurate RoI estimation can be affected by occlusions or long distances. To address these issues, we propose the Multi-Scale Grid Attention (MSGA) mechanism to explore multi-scale RoI analysis and emphasize the significance of different RoI regions. Furthermore, although existing methods treat depth estimation as a probabilistic process during training, they fail to fully leverage probabilistic properties during inference. We introduce a novel probabilistic post-processing method to enhance detection robustness. Experimental evaluations on the KITTI and Waymo datasets demonstrate that our approach achieves state-of-the-art performance.

We sincerely appreciate your reconsideration of our manuscript and look forward to the reviewers' feedback. For further correspondence, please contact Tingyu Zhang at the following address:

Institution: College of Computer Science and Technology, Jilin University, Changchun 130012, China
Telephone: +86 18013933973
Email: zhangty21@mails.jlu.edu.cn

Thank you for your attention to our manuscript.

Sincerely,
Tingyu Zhang