Detecting Hard Examples for Semantic Segmentation with Auxiliary 3D LiDAR

Yuriko Ueda†, Miho Adachi‡, and Ryusuke Miyamoto‡
†*Dept. of Computer Science, Graduate School of Science and Technology*
‡*Dept. of Computer Science, School of Science and Technology*
Meiji University, Kawasaki, Japan
Email:{ueda, miho, miya}@cs.meiji.ac.jp

***Abstract*—In this study, we propose a method for automatically detecting hard examples from unlabeled real-world data that lead to performance degradation in semantic segmentation models. The proposed method estimates obstacle regions using geometric information from a 3D LiDAR and generates obstacle masks by projecting them onto the image plane. These masks are compared pixel-wise with the semantic segmentation results, and a score based on the misclassification rate within obstacle regions is computed for each image. For evaluation, a semantic segmentation model trained on data collected at the Ikuta Campus of Meiji University was applied to unseen real-world data collected on the Tsukuba Challenge 2025 verification course. Experimental results confirmed that the proposed method could effectively identify images with many misclassified obstacle pixels.**

***Index Terms*—hard example, semantic segmentation, visual navigation, actual environment, LiDAR**

I. Introduction

Outdoor autonomous navigation and autonomous driving widely rely on sensing results by 3D LiDAR for environment perception. However, the high sensor cost of 3D LiDAR poses a barrier to its widespread deployment and practical implementation. To reduce the sensor cost, visual navigation based on semantic segmentation [1]–[6] has been proposed. Semantic segmentation is the task of assigning a class label to each pixel in an image. In this approach, drivable regions are detected from input images and used to enable autonomous navigation that combines road following with obstacle avoidance. As a result, the navigation performance strongly depends on the accuracy of semantic segmentation. For example, missed detections of obstacles may lead to overestimation of the drivable region, thereby increasing the risk of collision. In addition, inaccurate estimation of the boundary between drivable and non-drivable regions may cause the robot to deviate from the intended path and move outside the drivable area. Therefore, a highly accurate semantic segmentation model is essential for achieving safe and reliable autonomous navigation.

Previous work [7] reported that models trained on data collected from the target environment achieved better performance within that environment. However, semantic segmentation requires pixel-level annotations, and making ground-truth labels is highly labor-intensive. To reduce annotation costs, several studies have proposed semi-automatic dataset generation methods using 3D scanned data [8]–[12]. However, efficiently collecting training data that sufficiently reflects diverse real-world environments still remains a challenging problem.

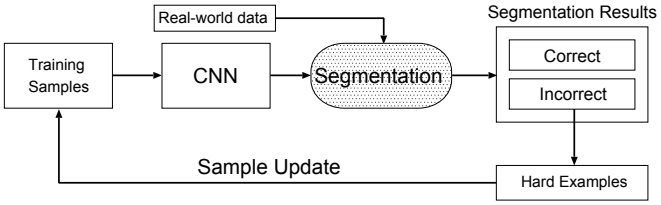


Fig. 1: Hard example mining to train a classifier for segmentation.

To address this challenge, we propose an automatic hard-example detection method toward a learning framework for semantic segmentation that focuses on hard examples. An overview of the learning framework is shown in Fig. 1. In this framework, hard examples are detected from newly acquired unlabeled data based on the inference results of a pre-trained semantic segmentation model, selectively annotated, added to the training dataset, and used for retraining. The proposed method is inspired by hard example mining, which has been widely used in object detection. In object detection, hard examples can be relatively easily defined because misdetections can be evaluated at the bounding-box level using training data with ground-truth labels. In contrast, semantic segmentation requires hard examples to be detected from newly acquired unlabeled data, making it difficult to define hard examples without ground-truth labels.

We introduce a 3D LiDAR during the data collection stage in addition to a monocular camera, and propose a method for detecting hard examples corresponding to obstacle classes using geometric information. Our previous work [13] attempted to detect hard examples using only image information; however, it was difficult to detect hard examples caused by obstacle classes that occupy small regions and exhibit significant dynamic changes. Therefore, the proposed method utilizes geometric information obtained from a 3D LiDAR, which is used only for hard-example detection. Specifically, obstacle masks generated from LiDAR point clouds are projected into image space, and the degradation in obstacle recognition performance is quantified based on the mismatch between these masks

and the semantic segmentation results. This enables automatic detection of images with a high proportion of misclassified pixels in the obstacle class from unlabeled real-world data.

This study is positioned as part of a broader effort to establish a training strategy based on hard example mining aimed at improving obstacle recognition performance in semantic segmentation. In particular, we focus on effectively incorporating misclassified pixels in obstacle regions into the training process, by investigating both the selection of hard examples and their application in learning. In this paper, as a first step, we propose a method that identifies hard examples based on the proportion of misclassified pixels within obstacle regions and prioritizes them during training. This approach allows misclassifications in obstacle regions to be efficiently reflected in the learning process, thereby improving obstacle detection performance. The main contributions of this study are summarized as follows:

- A method is proposed that leverages geometric information from LiDAR to automatically detect hard examples of the obstacle class from unlabeled real-world data.
- A metric is also proposed for identifying hard examples based on the proportion of misclassified pixels within obstacle regions.

II. Related Work

In this section, previous studies related to semantic-segmentation-based autonomous navigation and hard example detection are reviewed. First, LiDAR-based autonomous navigation and multimodal approaches are introduced. Next, studies on semantic-segmentation-based navigation and hard example mining are discussed. Finally, previous methods for hard example detection using only visual information are presented.

A. LiDAR-based and Multimodal Approaches

LiDAR-based navigation enables highly accurate environment perception using distance information and has been widely utilized for tasks such as obstacle avoidance and traversable area estimation [14]. In addition, multimodal approaches that integrate LiDAR and camera sensors have been reported to improve the robustness of obstacle detection and semantic region segmentation by combining geometric and visual information [15]. However, these methods require multiple sensors during inference, leading to increased system and computational costs.

In contrast, the proposed method does not use LiDAR during inference and instead utilizes it only as an auxiliary sensor for hard example detection. By performing the final navigation task using only a monocular camera, this study aims to realize a low-cost autonomous navigation system.

B. Visual Navigation based on Segmentation Results

Semantic segmentation-based navigation is a method for autonomous navigation that estimates traversable and obstacle regions. Since environment perception can be achieved using only a monocular camera, it has attracted considerable attention as a low-cost autonomous navigation approach.

The performance of semantic segmentation strongly depends on the training data, and previous studies have reported that training with data collected from the target environment is effective [7]. However, semantic segmentation requires pixel-level annotations, making dataset construction highly labor-intensive. Therefore, there is a need for methods that can efficiently train classifiers adapted to target environments under limited annotation costs.

C. Hard Example Mining for Semantic Segmentation

Hard example mining is a technique for improving classification performance by emphasizing samples that are likely to be misclassified. It has been widely applied to various tasks, including object detection. Representative methods include Online Hard Example Mining (OHEM) [16], which selects high-loss samples during training. Multimodal approaches have also been investigated for difficult-region detection and robust recognition [17]. Recent studies have further extended hard example mining strategies to representation learning and dense prediction tasks [18]. In addition, active learning approaches for semantic segmentation have been investigated to reduce annotation costs by selectively annotating informative samples from unlabeled data [19].

However, in semantic segmentation, pixel-level annotations are required, making it difficult to define hard examples at the image level. Furthermore, small obstacle regions and class imbalance often reduce the reliability of loss- or confidence-based sample selection methods. Moreover, datasets collected from target environments are often limited and unlabeled, making conventional loss-based approaches inapplicable. Therefore, alternative criteria for hard example detection are required.

D. Hard Example Detection using Only Visual Information

Previous work [13] proposed a method for detecting hard examples in semantic segmentation using sequential images. The method detects performance degradation by measuring changes in class areas between consecutive frames.

While the method was effective for ground classes such as roads and sidewalks, it was less effective for obstacle classes because their regions are typically small and dynamically changing. These limitations indicate that area-based metrics alone are insufficient for stable hard example detection in obstacle classes.

To address this issue, this study introduces a 3D LiDAR as an auxiliary sensor for estimating obstacle regions geometrically. By integrating point cloud and image data, the proposed method enables more reliable detection of hard examples in obstacle classes from unlabeled real-world data.

III. Hard Example Detection Using Auxiliary 3D LiDAR

In this section, we propose a method for detecting hard examples including obstacle class using LiDAR information.

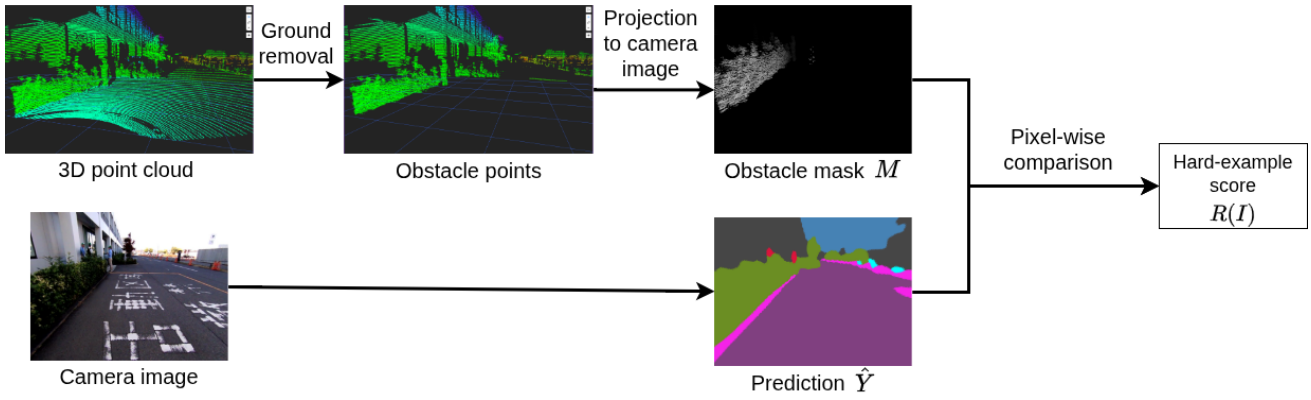


Fig. 2: Overview of the proposed method for hard example detection.

First, the overall framework and problem formulation are introduced, followed by the LiDAR-based obstacle mask generation method and the hard-example detection process.

A. Overview of Proposed Framework

Figure 2 illustrates the overall workflow of the proposed method. First, a robot equipped with a monocular camera and a 3D LiDAR simultaneously acquires image and point cloud data. Next, the semantic segmentation model generates a prediction map $\hat{Y}$ from the input image I. Meanwhile, the point cloud data are processed to generate an obstacle mask M. Finally, $\hat{Y}$ and M are compared to compute the misclassification rate $R(I)$ within obstacle regions. Hard examples are detected based on the score $R(I)$.

B. Problem Formulation and Definition of Hard Examples

Let the input image, semantic segmentation result, and LiDAR-based obstacle mask be denoted by I, $\hat{Y}$, and M, respectively. The pixel i in M_i is defined as

$$M_i = \begin{cases} 1 & \text{if pixel } i \text{ is estimated as an obstacle by LiDAR} \\ 0 & \text{otherwise} \end{cases}.$$

In this study, a pixel i is considered misclassified if

$$M_i = 1 \quad \text{and} \quad \hat{Y}_i \notin C_{obs}, \tag{1}$$

where C_{obs} denotes the set of obstacle classes. That is, the pixel is labeled as an obstacle by LiDAR but predicted as a non-obstacle by the segmentation model.

The total number of obstacle pixels and the number of misclassified pixels are defined as

$$N_{obs}(I) = \sum_i \mathbf{1}\left(M_i = 1\right), \text{and} \tag{2}$$

$$N_{err}(I) = \sum_i \mathbf{1}\left(M_i = 1 \wedge \hat{Y}_i \notin C_{obs}\right). \tag{3}$$

Thus, the misclassification rate is computed as

$$R(I) = \frac{N_{err}(I)}{N_{obs}(I)}. \tag{4}$$

An input image I is regarded as a hard example for the obstacle class if $R(I) \geq \tau$. Here, images with $N_{obs}(I) = 0$ are excluded from consideration.

This approach enables the automatic identification of samples that degrade the performance of the semantic segmentation model from real-world data, without requiring manual annotation. A high value of $R(I)$ indicates that many pixels estimated as obstacles by LiDAR are predicted as non-obstacles by the segmentation model. Therefore, the proposed score can be used as an indicator of obstacle-recognition failure in unlabeled environments.

C. Obstacle Mask Generation from LiDAR Output

The point cloud acquired by the 3D LiDAR is first processed to remove ground points, and the remaining points are treated as obstacle points. Patchwork++ [20] is used for ground removal because it enables robust ground detection in environments containing slopes and steps. Next, a forward-distance range filter is applied to the obstacle points to remove distant points and long-range noise. This focuses the detection on regions that directly affect the robot's navigation.

The resulting obstacle points are then projected onto the image plane using the extrinsic parameters between the LiDAR and camera, as well as the camera intrinsic parameters. After projection, a binary mask M is generated by setting the corresponding pixels to 1, which is used as the obstacle mask.

IV. Evaluation

In this section, the effectiveness of the proposed method is evaluated using real-world datasets collected in outdoor environments. First, the experimental setup and evaluation metrics are described, followed by the evaluation results and discussion.

A. Experimental Setup

The semantic segmentation model used in this study was trained using data collected on the Ikuta Campus of Meiji University. For evaluation, ROS bags recorded on the Tsukuba Challenge 2025 verification course were used as test data. Since the Tsukuba environment differs from the training

environment, the evaluation allows assessment of the proposed method under unseen real-world conditions.

Although datasets such as Cityscapes could also be used, their environments differ significantly from those targeted in this study, resulting in an excessively large domain gap. Therefore, in this study, evaluation was conducted using data collected from different locations within outdoor environments in Japan.

B. Dataset

Data were collected in outdoor environments using a robot equipped with a monocular RGB camera (640 × 480, approximately 30 fps) and a 3D LiDAR (Lslidar CH256X [21], approximately 10 Hz). All images were undistorted using calibrated camera parameters before LiDAR point projection and evaluation. Example raw images and ground-truth labels from each environment are shown in Fig. 3.

The dataset consists of eight classes: Road, Sidewalk, Terrain, Building, Static, Dynamic, Sky, and Vegetation. Since these are functional categories for navigation, unseen objects can generally be represented by existing classes such as Static or Dynamic. Building, Static, Dynamic, and Vegetation are treated as obstacle classes. The semantic segmentation model used in this study was PSPNet [22] with a ResNeXt101d [23] backbone. The model was trained on the Ikuta dataset and applied to the Tsukuba dataset without additional fine-tuning.

LiDAR point clouds were synchronized using timestamps and projected onto the image plane using calibrated LiDAR–camera extrinsics and camera intrinsics. The extrinsic parameters were manually fine-tuned for each scene to align projected obstacle points with image obstacles. When generating obstacle regions, Patchwork++ and a forward-distance range filter were applied. The parameters were preliminarily adjusted to balance the suppression of ground misdetections and long-range noise.

For evaluation, multiple scene segments were extracted from the recorded ROS bags. The resulting test dataset consists of the following scenes:

- Scene 1: A scene containing metallic obstacles that were frequently missed in previous experiments, consisting of 83 frames.
- Scene 2: A scene near a park crosswalk with frequent dynamic obstacles (pedestrians and bicycles) and static obstacles (e.g., chairs). Two frames affected by robot vibrations when crossing the sidewalk–road step were excluded, consisting of 51 frames.
- Scene 3: A scene from the front of Kenkyu-Gakuen Station toward the station entrance, containing small obstacles, people, walls, and glass surfaces, consisting of 70 frames.

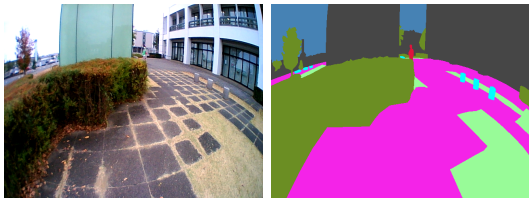
C. Evaluation Metrics

As an evaluation metric, the proportion of pixels within LiDAR-estimated obstacle regions that are incorrectly predicted as non-obstacles, denoted by R, is used. Let N_{obs} be the total number of pixels within obstacle regions, and N_{err} be the number of pixels incorrectly predicted as non-obstacles. Then, R is defined as $R = N_{err}/N_{obs}$. For evaluation on real-world data, the inference results were visually inspected, and each image was manually labeled with a hard example label $y(I) \in \{0, 1\}$. Here, $y(I) = 1$ indicates that the image was identified as a hard example by human inspection.

By sweeping the threshold τ, the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are obtained, and the true positive rate (TPR) and false positive rate (FPR) are calculated as

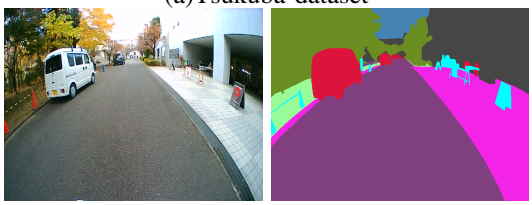
$$\mathrm{TPR} = \frac{TP}{TP + FN}, \quad \mathrm{FPR} = \frac{FP}{FP + TN}. \tag{5}$$

From these values, the ROC curve (TPR–FPR) is constructed, and the area under the curve (AUC-ROC) is used as the evaluation metric. Since the proposed method depends on the threshold τ, the threshold is swept to evaluate the trade-off between true positive and false positive detections. Images with $N_{obs} = 0$ are excluded from evaluation to remain consistent with the definition of the proposed method.



RGB image Segmentation GT

(a)Tsukuba dataset



RGB image Segmentation GT

(b)Ikuta dataset

Fig. 3: Examples of the Ikuta dataset used for training and the Tsukuba dataset used for evaluation.

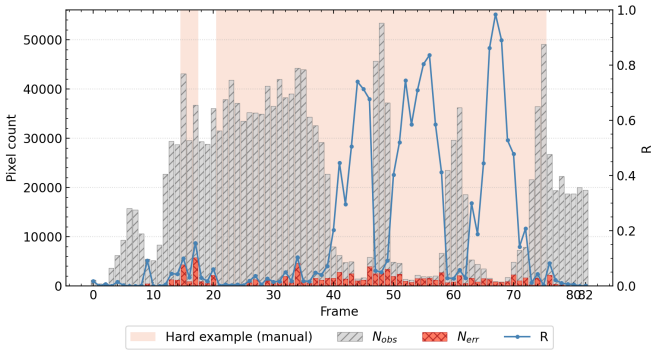


Fig. 4: Evaluation results for Scene 1.

D. Results

The per-frame values of N_{obs}, N_{err}, and the score $R(I)$ for each test dataset are shown in Figs. 4–6. The corresponding

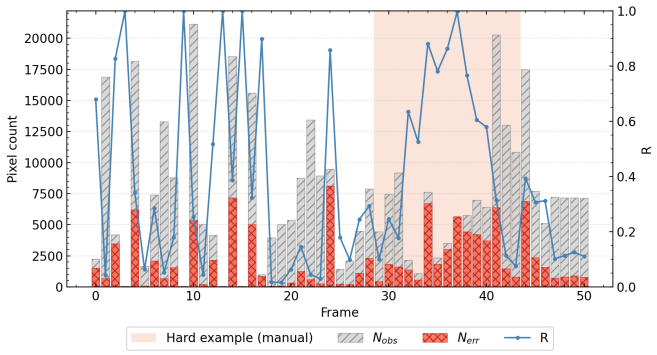


Fig. 5: Evaluation results for Scene 2.

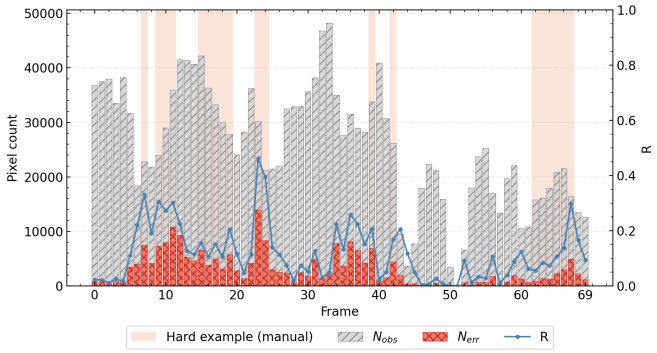


Fig. 6: Evaluation results for Scene 3.

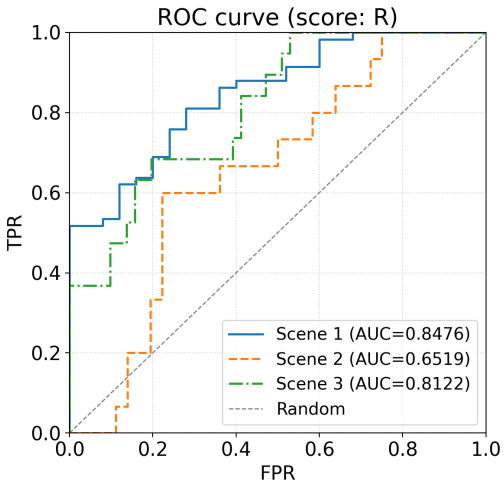


Fig. 7: ROC curves and AUC-ROC values for each evaluation scene.

RGB image | Segmentation result | LiDAR-based obstacle mask

Fig. 8: Example image from Scene 1 labeled as a hard example ($R = 0.01$).

ROC curve and AUC-ROC are presented in Fig. 7.

In Scene 1, as shown in Fig. 4, images with scores R greater than or equal to 0.2 are all classified as hard examples. However, some images with $R < 0.2$ are also identified as hard examples. An example is shown in Fig. 8, where ground regions are misclassified as obstacles.

In Scene 2, as shown in Fig. 5, images classified as hard examples tend to exhibit relatively high scores R. However, among frames 5–22, some images not labeled as hard examples still have scores near 1.0. Indeed, as illustrated in Fig. 9, most of these images contain very few obstacles.

In Scene 3, as shown in Fig. 6, the maximum score R for images classified as hard examples is generally around 0.5, which is lower than in the other two scenes. An example of a hard example is frame 23, shown in Fig. 10. In this frame, parts of the building on the left and the pole on the right are misclassified as non-obstacles.

Finally, the ROC curve is shown in Fig. 7. The AUC-ROC exceeds 0.8 in scenes with relatively low noise levels, while in scenes containing significant noise, it remains around 0.6. These results indicate that the proposed score R can discriminate hard examples from normal samples to some extent.

E. Discussion

First, in Scene 1, some hard examples were not detected even though they affect navigation. As shown in Fig. 8, these cases included situations where ground regions were misclassified as obstacles. Since the score R only considered misclassifications within obstacle regions, errors in ground regions were not reflected in the score. Consequently, some hard examples remained undetected.

Second, in Scene 2, several frames exhibited scores close to 1.0 despite not being labeled as hard examples. This occurred because N_{obs} was very small, such that even a small number of misclassified pixels resulted in a high value of R. In many of these frames, LiDAR point cloud noise was detected as obstacles, artificially increasing the score. Therefore, LiDAR noise could reduce the reliability of hard example detection.

Third, in Scene 3, obstacles made of glass were not detected by LiDAR. Since LiDAR beams could pass through glass surfaces, these regions were not included in the obstacle mask used to compute R. As a result, even when such regions were misclassified, the score remained relatively low. This suggested that obstacles that were difficult to detect using LiDAR could reduce the effectiveness of the proposed method.

Overall, these observations indicate that the performance of the proposed method is affected by several factors, including ground-region misclassification, LiDAR noise, and obstacles with poor LiDAR detectability. Possible improvements include incorporating image-based cues together with LiDAR information, applying temporal filtering to suppress transient point cloud noise, and introducing normalization or obstacle-size constraints to reduce the influence of sparse LiDAR detections.

In addition, the current evaluation was conducted using

RGB image Segmentation result LiDAR-based obstacle mask

Fig. 9: Example image from Scene 2 that is not labeled as a hard example despite $R = 1.00$.

RGB image Segmentation result LiDAR-based obstacle mask

Fig. 10: Example image from Scene 3 labeled as a hard example ($R = 0.46$).

multiple scenes from the Tsukuba Challenge course only. Therefore, further evaluation on diverse environments and datasets will be necessary to investigate the robustness and generalization capability of the proposed method.

V. Conclusion

In this study, we proposed a method for automatically detecting hard examples in obstacle classes for semantic segmentation using geometric information obtained from a 3D LiDAR in addition to a monocular camera. The proposed method uses the misclassification rate within LiDAR-based obstacle regions as a score for hard-example detection.

For evaluation, a model trained on data collected at the Ikuta Campus of Meiji University was applied to unseen real-world data collected on the Tsukuba Challenge 2025 course. Experimental results demonstrated that the proposed method could detect hard examples from unlabeled real-world data. Furthermore, ROC curves and AUC-ROC confirmed the effectiveness of the proposed score.

Several limitations, including ground-region misclassification and LiDAR noise, were also identified. Future work includes improving hard-example detection accuracy and investigating learning strategies that utilize the detected hard examples.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP25K01236.

References

[1] R. Miyamoto, M. Adachi, H. Ishida, T. Watanabe, K. Matsutani, H. Komatsuzaki, S. Sakata, R. Yokota, , and S. Kobayashi, "Visual navigation based on semantic segmentation using only a monocular camera as an external sensor," *J. Robot. Mechatron.*, vol. 32, no. 6, pp. 1137–1153, 2020.

[2] M. Adachi, K. Honda, J. Xue, H. Sudo, Y. Ueda, Y. Yuda, M. Wada, and R. Miyamoto, "Practical implementation of visual navigation based on semantic segmentation for human-centric environments," *J. Robot. Mechatron.*, vol. 35, no. 6, pp. 1419–1434, 2023.

[3] R. Miyamoto, Y. Nakamura, M. Adachi, T. Nakajima, H. Ishida, K. Kojima, R. Aoki, T. Oki, and S. Kobayashi, "Vision-based road-following using results of semantic segmentation for autonomous navigation," in *Proc. ICCE Berlin*, 2019, pp. 194–199.

[4] M. Adachi, S. Shatari, and R. Miyamoto, "Visual navigation using a webcam based on semantic segmentation for indoor robots," in *Proc. SITIS*, 2019.

[5] M. Adachi, K. Honda, and R. Miyamoto, "Turning at intersections using virtual lidar signals obtained from a segmentation result," *Journal of Robotics and Mechatronics*, vol. 35, no. 2, pp. 347–361, 2023.

[6] Y. Ueda, M. Adachi, J. Morioka, M. Wada, and R. Miyamoto, "Data augmentation for semantic segmentation using a real image dataset captured around the tsukuba city hall," *Journal of Robotics and Mechatronics*, vol. 35, no. 6, pp. 1450–1459, 2023.

[7] R. Miyamoto, M. Adachi, Y. Nakamura, T. Nakajima, H. Ishida, and S. Kobayashi, "Accuracy improvement of semantic segmentation using appropriate datasets for robot navigation," in *Proc. CoDIT*, 2019, pp. 1610–1615.

[8] M. Wada, M. Adachi, Y. Ueda, and R. Miyamoto, "Dataset genreratoin for semantic segmentation from 3d scanned data considering domain gap," in *Proc. CoDIT*, 2023, pp. 1711–1716.

[9] M. Wada, M. Adachi, and R. Miyamoto, "Effect of varied datasets on training of a segmentation model used in visual navigation," in *IEEE Int. Conf. Big Data*, 2023, pp. 4350–4355.

[10] M. Wada, H. Sudo, M. Adachi, and R. Miyamoto, "Does a dense point cloud for training data generation improve segmentation accuracy?" in *IEEE Int. Conf. Big Data*, 2023, pp. 4356–4361.

[11] M. Wada, Y. Ueda, J. Morioka, M. Adachi, and R. Miyamoto, "Dataset creation for semantic segmentation using colored point clouds considering shadows on traversable area," *J. Robot. Mechatron.*, vol. 35, pp. 1406–1418, 12 2023.

[12] M. Wada, Y. Ueda, M. Adachi, and R. Miyamoto, "Area-wise augmentation on segmentation datasets from 3d scanned data used for visual navigation," in *Proc. CoDIT*, 2024, pp. 1499–1504.

[13] Y. Ueda, M. Wada, M. Adachi, and R. Miyamoto, "Hard example detection at actual environment for semantic segmentation used in visual navigation," in *Proc. World Congress on Mechanical, Chemical, and Material Engineering*, Aug 2024.

[14] T. Ort, L. Paull, and D. Rus, "Autonomous vehicle navigation in rural environments without detailed prior maps," in *Proc. IEEE Int. Conf. on Robotics and Automation*, May 2018, pp. 2040–2047.

[15] K. Tsiakas, D. Alexiou, D. Giakoumis, A. Gasteratos, and D. Tzovaras, "Leveraging multimodal sensing and topometric mapping for human-like autonomous navigation in complex environments," in *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, 2023, pp. 7415–7421.

[16] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[17] D. Biswas, A. Scouten, H. Gong, F. Wang, and J. Tešić, "Mopac+: Multimodal modeling of the pavement cracks using hard-example mining and context-aware feature aggregation," *Measurement*, vol. 272, p. 120896, 05 2026.

[18] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang, "Hard patches mining for masked image modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10 375–10 385.

[19] N. Qiao, Y. Sun, C. Liu, L. Xia, J. Luo, K. Zhang, and C.-H. Kuo, "Human-in-the-loop video semantic segmentation auto-annotation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5870–5880.

[20] S. Lee, H. Lim, and H. Myung, "Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud," in *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, 2022, pp. 13 276–13 283.

[21] "Ls lidar," www.lslidar.com.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[23] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.