Evaluating SAM-Based Labeling Approaches for Autonomous Driving in Korean Traffic Scenarios

Young-Woong Jun[1], Sang-Chul Kim[2]
Department of Computer Science
Kookmin University, Seoul, South Korea
Email: {jun0woong, sckim7}@kookmin.ac.kr

Abstract—This study evaluates the performance of autonomous driving models on the 42dot dataset, which represents the complex road environments in Korea. The dataset was labeled using both automatic and semiautomatic techniques with the segment anything model (SAM). Fine-tuning a DeepLabv3+ model pretrained on the Cityscapes dataset using the labeled data enabled comparative analysis between the two datasets. The experimental results demonstrated that the model trained on the SAM-labeled 42dot dataset outperformed the Cityscapes-trained model in handling complex road scenarios despite minor class imbalance challenges. In addition, SAM-based automatic labeling demonstrated higher efficiency and consistency than the semiautomatic approach, emphasizing the potential for scalable annotation workflows. These findings highlight the importance of using domain-specific datasets and automated labeling techniques to enhance the effectiveness of autonomous driving systems.

Index Terms—Segment Anything Model, 42dot Dataset, DeepLabv3+, Autonomous Driving, Semantic Segmentation

I. Introduction

The rapid advancement of autonomous driving technologies has increased the demand for high-quality datasets. Although object detection techniques have significantly improved, accurately recognizing objects in complex road environments remains a challenge. The Cityscapes dataset includes various traffic scenarios; however, it falls short of capturing the intricacies and diversity of road conditions in specific regions, such as Korea.

The 42dot dataset collected from real-world Korean roads provides a more comprehensive representation of the high object density and diverse traffic situations that are prevalent in such environments. The dataset is well-suited for training autonomous driving models to handle complex urban scenarios. The availability of diverse object classes makes this dataset particularly useful for addressing the challenges posed by dense traffic and diverse road conditions.

In this study, we applied the segment anything model (SAM) to perform both automatic and semiautomatic labeling using the 42dot dataset. The SAM, which was originally designed to segment all objects in an image, has been adapted to label specific objects using bounding box coordinates as prompts. This modification improves the accuracy and consistency of the labeling process, particularly for complex road scenes.

We fine-tuned a DeepLabv3+ model pretrained on the Cityscapes dataset using the labeled 42dot data. This allowed us to compare the performance of the model on both datasets and assess the impact of domain-specific datasets. The results of this study highlight the benefits of using customized datasets to improve the real-world applicability of autonomous driving systems.

The primary objectives of this study are threefold:

- To evaluate the performance of autonomous driving models trained on the 42dot dataset using SAM-based automatic and semiautomatic labeling techniques.
- To compare the performance of models fine-tuned on the 42dot dataset with models trained on the Cityscapes dataset.
- To demonstrate the effectiveness of automated labeling and domain-specific datasets in improving model performance in complex urban scenarios.

This study contributes to the field in several ways, including the following:

- We introduce an efficient labeling pipeline using the SAM that incorporates bounding box coordinates as prompts to enhance the precision of segmentation masks.
- We perform comparative analysis of the 42dot and Cityscapes datasets to highlight the importance of dataset customization for specific environments.
- We demonstrate that SAM-based automatic labeling, reinforced with bounding box prompts, provides higher consistency and efficiency than semiautomatic methods, thereby contributing to the development of more scalable annotation workflows.

The remainder of this paper is organized as follows: Section II provides an overview of the datasets and describes the DeepLabv3+ model architecture and the fine-tuning process. In Section III, we compare the labeling techniques and model performance across the datasets. Finally, Section IV concludes the study and proposes future research directions.

II. Methodology

A. Dataset Preparation

The 42dot dataset [4] comprises 13,134 high-resolution images that reflect the complex road environments in Korea.

Camera Setup: Images were captured from three cameras: front-center (60° field of view), front-left, and front-right (120° field of view each). - Image Resolution: 1,920 × 1,208 pixels.

Object Classes: The dataset includes seven major object classes (vehicles, pedestrians, two-wheelers, etc.) to capture diverse traffic conditions with high object density.

The dataset is designed to improve the performance of object recognition models in challenging road environments with complex traffic scenarios.

B. Automatic and SemiAutomatic Labeling

1) Automatic Labeling using SAM: In this study, we applied the SAM [2] to automatic labeling of the 42dot dataset. Although the original SAM labels all objects in an image, we incorporated bounding box coordinates as prompts to label only specific objects. This approach generated 13,134 images with the corresponding mask images.

Advantages of Automatic Labeling: Using bounding boxes, labeling consistency and accuracy were improved; [3] the labeling process also became more efficient, thereby reducing both time and cost.

2) SemiAutomatic Labeling using CVAT: We also employed semiautomatic labeling using the SAM integrated into the CVAT tool. In this process, the user manually selects the object of interest, and the SAM generates a precise segmentation along the boundaries of the object. Thus, we created 5,289 images with corresponding mask images. Fig. 1 shows the structure of the SAM.

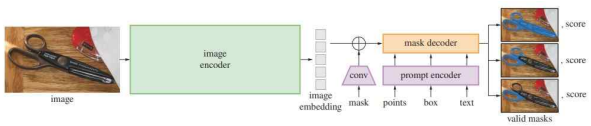


Fig. 1. Structure of the SAM using bounding box coordinates as prompts.

C. DeepLabv3+ Model

The DeepLabv3+ model [1] is a powerful image segmentation model that is designed to handle objects of various sizes with high precision.

Atrous Spatial Pyramid Pooling (ASPP): ASPP integrates context information from multiple resolutions, thereby enabling accurate segmentation even in complex scenes.

Decoder Module: The decoder module enhances the segmentation of object boundaries, thereby improving recognition performance in challenging road environments. In this study, we fine-tuned a DeepLabv3+ model pretrained on the Cityscapes dataset using the 42dot dataset to improve its performance. Fig. 2 presents the architecture of the DeepLabv3+ model.

III. Evaluation

A. Comparison of Labeling Techniques

The experimental results showed no significant difference in mean intersection over union (mIoU) values between the datasets constructed using the automatic and semiautomatic labeling techniques. This implies that the automatic labeling technique (SAM) is efficient and exhibits performance comparable to that of semiautomatic labeling. Therefore, using automatic labeling techniques to construct large-scale datasets can save time and cost in the development of autonomous driving systems. Fig. 3 shows examples of datasets labeled automatically and semiautomatically.

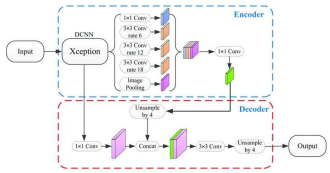
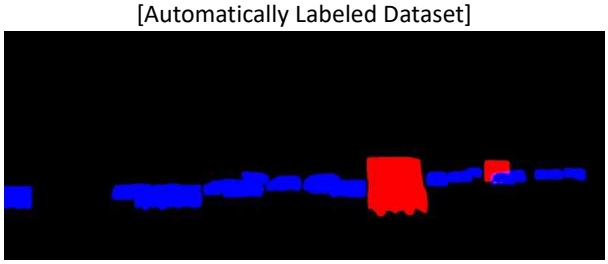


Fig. 2. Structure of DeepLabv3+.

[Automatically Labeled Dataset]



[SemiAutomatically Labeled Dataset]

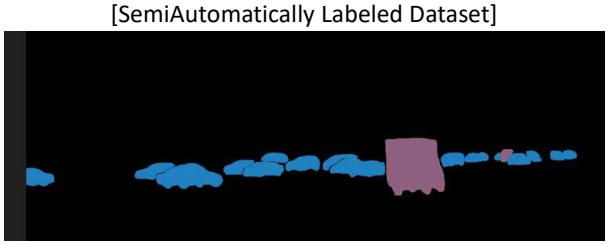


Fig. 3. Examples of automatically and semi-automatically labeled datasets.

B. Model Performance Comparison

Compared with the DeepLabv3+ model trained on the Cityscapes dataset, the model fine-tuned with the 42dot dataset exhibited lower mIoU values. This was likely due to class imbalance in the 42dot dataset. Specifically, the 42dot dataset-fine-tuned model may not have sufficiently learned certain classes because the data for specific classes were relatively scarce or various objects were grouped into the same class. Fig. 4 compares mIoU values across datasets.

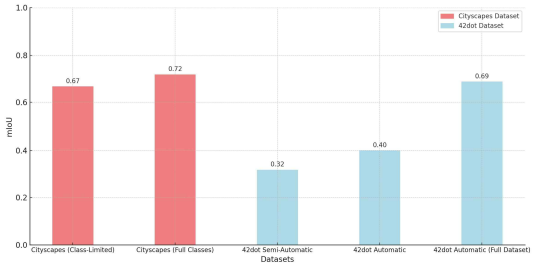


Fig. 4. Comparison of mIoU values across datasets and labeling techniques.

In addition, retraining the model by reducing the number of classes in the Cityscapes dataset resulted in lower mIoU values than those obtained by the model trained with the original number of classes. This result indicates that the diversity of the number of classes in datasets affects model performance.

IV. Conclusion

In this study, we analyzed the efficiency of SAM-based automatic labeling and evaluated the impact of SAM-labeled data on the performance of autonomous driving models. The experimental results demonstrated that the model trained on the SAM-labeled data performed well in complex road scenarios and exhibited higher consistency and time-saving advantages than that trained on semiautomatically labeled data. Specifically, the DeepLabv3+ model trained on the class-refined Cityscapes dataset achieved a mIoU value of 0.67, whereas the model trained on the automatically SAM-labeled 42dot dataset achieved a mIoU value of 0.69, demonstrating performance improvement. However, there were class imbalance issues with the SAM-labeled 42dot dataset, particularly for classes such as pedestrians and two-wheelers. To address these issues, we applied class-specific weights (1.0, 1.5, and 2.0) to the loss function during training. Although this fixed-weight strategy improved model performance, alternative methods such as dynamic weighting schemes or data augmentation could further improve model performance.

This study highlights the potential of using Korea-specific datasets to develop autonomous driving technologies and demonstrates the utility of SAM-based labeling. Future research will focus on more advanced solutions to effectively address class imbalance. In addition, performance evaluations in real-world environments or simulators will be conducted to further validate and improve the results of this study.

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “EncoderDecoder with Atrous Separable Convolution for Semantic Image Segmentation,” *arXiv preprint arXiv:1802.02611*, 2018.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, P. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollar, and R. Girshick, “Segment Anything,” *arXiv preprint arXiv:2304.02643*, 2023.

[3] B. Cheng, Y. Tai, S. Li, X. Wan, C. Fu, X. Liang, C. Chen, H. Li, and X. Tong, “Fully Convolutional Networks for Panoptic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6224-6237, June 2023, doi: 10.1109/TPAMI.2022.3172947.

[4] “Multi-Camera Multi-Object Tracking Dataset for Autonomous Driving,” [Online]. Available: https://42dot.ai/openDataset/ad/mcmot. [Accessed: Oct. 16, 2024].