HaWANet: Road Scene Understanding with Multi-modal Sensor Data using Height-Width-driven Attention Network

Soumick Chatterjee[1,2,3][0000−0001−6869−4019], Jiahua Xu*[1,2,4][0000−0003−3672−373X], Adarsh Kuzhipathalil*[1,5], and Andreas Nürnberger[1,2][0000−0003−4311−0624]

[1] Faculty of Computer Science, Otto von Guericke University Magdeburg, Germany
soumick.chatterjee,andreas.nuernberger@ovgu.de
[2] Data and Knowledge Engineering Group, Otto von Guericke University Magdeburg, Germany
[3] Genomics Research Centre, Human Technopole, Milan, Italy
soumick.chatterjee@fht.org
[4] Hertie Institute for Clinical Brain Research, Tuebingen, Germany
Jiahua.Xu@med.uni-tuebingen.de
[5] XenomatiX NV, Leuven, Belgium
adarsh.kuzhipathalil@xenomatix.com

Abstract. In recent years, the field of autonomous vehicles and driverless technology has seen remarkable advancements, driven by contributions from mainstream automotive manufacturers and open-source projects. This research aims to develop a pipeline for road scene understanding through semantic segmentation. The proposed pipeline utilises a multi-modal segmentation model, incorporating greyscale images and point cloud data from Xenolidar, specifically designed to capture the structural priors of highway road scenes. The fusion of input modalities and the design of an encoder-decoder architecture with a novel attention scheme called HaWANet is introduced, which focuses on the height and width contextual information to improve the accuracy of road segmentation, are the primary aspects explored for the proposed model. The output of the encoder is a two-dimensional point cloud, which effectively represents the road's planar nature, and is crucial for improving the accuracy of road segmentation, particularly in edge cases, addressing current challenges in autonomous driving research. This research, aimed at addressing the segmentation problem for multimodal sensor data, has presented significant performance improvement over single-modal approaches.

Keywords: Road Scene Understanding · Multimodal Segmentation · Deep Learning · Attention.

* equal contribution

1 Introduction

Recent years have seen significant advances in the field of autonomous vehicles and driverless technology, with mainstream automotive manufacturers and open-source projects contributing high-quality research to solve autonomous driving problems. This progress can be attributed to the availability of open-source datasets and improvements in computational capabilities. This research aims to develop a pipeline for road scene understanding through semantic segmentation.

Despite significant improvements, fully autonomous driving remains a challenging goal due to its safety-critical nature. Current research focuses on addressing edge cases and improving system robustness. The choice of sensors in the perception stack is crucial in this context. LIDAR (Light Detection and Ranging) is widely used for long-range obstacle detection and is a key component of autonomous vehicle perception systems. It functions by emitting light pulses and measuring the time it takes for them to return after bouncing off objects, generating extensive data.

The challenge lies in extracting usable information from these data to perceive the vehicle's surroundings. Advances in LiDAR technology now provide high-quality 3D information. Combining 3D LiDAR data with the rich semantic information from camera images can enhance perception algorithms. Such data fusion strategies leverage the strengths of each sensor modality. For example, LiDAR performs well in low-light conditions, whereas cameras provide semantically rich data.

The efficient fusion of sensor modalities and the design of neural networks to solve various computer vision problems are crucial to the advancement of autonomous driving. This research explores data fusion between multimodal sensor data from a solid-state LiDAR with a proposed segmentation model. The aim is to contribute to improving robustness and accuracy, particularly in edge cases, by addressing current challenges in autonomous driving research.

1.1 Background

Perception systems form a crucial component of the autonomous driving stack. Over the past decade, these systems have significantly evolved, integrating high-accuracy sensor systems to process data about the vehicle's surroundings. Modern driverless cars use a sensor stack that includes cameras, LiDARs (light detection and range), radars (radio detection and range), and ultrasonic sensors. Real-time processing of data from these multimodal sensors presents a major challenge in autonomous driving. This research focuses on LiDARs, whose capabilities in low-light and nighttime conditions are invaluable, despite ongoing debates about their necessity in autonomous vehicles.

LiDARs in Automotive Perception Systems LiDAR sensors have become a standard component in the perception stack for autonomous vehicles. Although there is a debate about their necessity, LiDARs provide efficient and accurate

3D perception with minimal post-processing. They offer precise 360-degree vision and faster depth sensing compared to other methods such as stereo vision. The requirements for LiDAR sensors in safety-critical applications such as autonomous vehicles include long range, real-time response, high spatial resolution, and tolerance to sunlight [8].

LiDAR technology works on the Time of Flight (ToF) principle, which measures the time between sending and receiving reflection of a light beam [1]. LiDARs generate point clouds by repeating these point measurements, and they are classified into two main types: Spinning LiDARs and Solid-state LiDARs.

Spinning LiDARs feature a rotating element that scans light around it and a receiver element that calculates the ToF to generate point clouds. These were the first 360-degree scanning devices used in the autonomous vehicle industry.

Solid-state LiDARs, a term from the semiconductor industry, use static scanners and receivers. Notable types include Microelectromechanical Systems (MEMS) based LiDARs, which use tiny mirrors to control the direction of the laser beam by adjusting the tilt angle with a stimulus voltage [15]. Another type is Vertical-Cavity Surface-Emitting Lasers (VCSEL) scanners, known for their precision and efficiency [2,16]. The LiDAR scanner used in this research employs VCSEL technology [17].

HANet: Attention for Semantic Segmentation Incorporating the intrinsic nature of different scenes into computer vision tasks is an area with relatively little research. Specifically, for road scene understanding, many inherent presets can enhance algorithm design. Height-driven Attention Networks (HANet), proposed by Sungha Choi et al. [3], exemplifies this approach.

HANet capitalises on the structural priors of road scenes captured by front-mounted cameras in vehicles. This approach splits road scene images into three height-based regions and evaluates the class probability distribution in these regions compared to the whole image. The study found lower entropy in the height-based regions, confirming the potential to integrate these structural priors into semantic segmentation frameworks. HANet adds a height-driven attention mechanism to the segmentation framework, leveraging these spatial distributions to improve scene understanding.

Deep Data Fusion When multimodal data is available, as in this research, a fusion strategy can be used to potentially improve the performance of the semantic segmentation problem. In general, data fusion can be categorised as: early fusion, late fusion, and hybrid fusion [19] - presented in Fig. 1

Early fusion combines different modalities before feeding them into feature extraction models. Camille Couprie et al.[5] introduced early fusion in 2013 for indoor scene segmentation by combining RGB and depth information using a Laplacian pyramid. Another notable method, FuseNet [6], fuses RGB and depth modalities at each feature extraction level within two encoder networks.

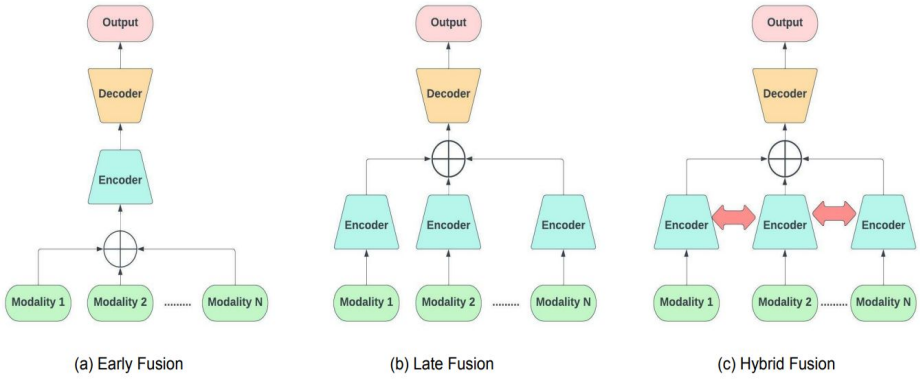

Fig. 1. Deep Multimodal fusion strategies

Late fusion involves feeding each modality into separate encoder models and combining the extracted features at a later stage. Gupta et al.[4] employed late fusion in 2014 by extracting RGB and depth features with two encoders and combining them using an SVM classifier. Valada et al.[13] later summed features from different modalities and fed this joint representation into a series of convolution layers.

Hybrid Fusion combines early and late fusion techniques to enhance segmentation quality. Valada et al.[12] developed a Self-Supervised Model Adaptation (SSMA) module, which adapts semantically mature feature representations at different scales. The SSMA blocks successfully exploit modal-specific features and enhance discriminative factors in the feature map.

2 Methodology

2.1 Dataset

This research uses data from Xenolidar, consisting of two primary modalities: a 2D greyscale image and a 3D point cloud. The initial phase of this research involved data preparation and annotation, which is recognised as the most time-consuming and costly aspect of machine learning projects. This research utilised advanced deep learning methods to minimise the required time and effort.

Data Collection The data collection setup comprises Xenolidar housings developed by Xenomatix and an RGB camera that captures the same scene. Data from both sensors are time-stamped. The RGB camera is included to generate accurate predictions from existing model architectures trained on RGB images.

Generating Annotations Initial attempts to generate rough segmentation masks involved inputting Xenolidar greyscale images into a pretrained segmentation network. Two methods were tested: stacking the greyscale image to create three channels and modifying the network for single-channel input. Both methods were unsuccessful due to the unique nature of Xenolidar images, leading to an alternative approach.

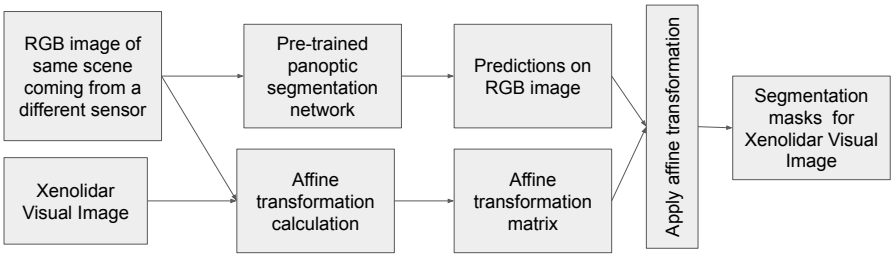

Fig. 2. Designed data preparation scheme where the affine transformation between RGB images and greyscale images from two different sensors is calculated first and then used to transfer segmentation masks

The data flow of the annotation scheme, illustrated in Figure 2, starts with RGB images processed by the DeepLabV3 model, pretrained on the Cityscapes dataset. The semantic segmentation masks obtained for RGB images are transferred to the Xenolidar image space using affine transformations. For this, three corresponding points are selected from a sample RGB and greyscale Xenolidar image. Although manual selection is possible, automation using SIFT feature correspondence is preferred for accuracy. From the list of corresponding points, three distant points are chosen to calculate the affine transformation matrix, which is used to transform RGB images into the Xenolidar image space. This process is repeated multiple times to refine the transformation matrix.

The same affine transformation matrix is applied to convert segmentation masks of RGB images into masks for Xenolidar greyscale images, creating a rough set of segmentation masks. These masks are manually corrected using a modified Labelme tool [14], which accepts Xenolidar greyscale images and displays the corresponding RGB images and point-cloud overlays. Nine classes from the Cityscapes dataset were selected for annotation: car, truck, person, road, sidewalk, building, sky, vegetation, and bicycle + motorcycle.

The dataset, consisting of 12,000 frames (8,000 recorded in Belgium and 4,000 in Japan), required over four months of manual correction despite automation to achieve the desired quality.

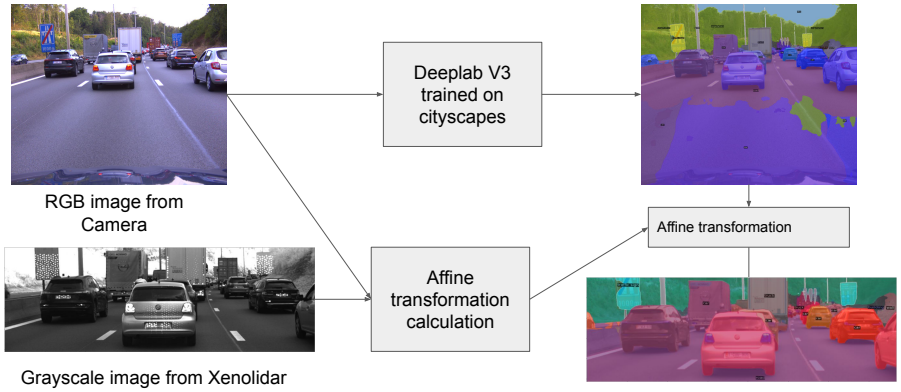

Fig. 3. Sample result from the data annotation pipeline

2.2 Semantic Segmentation Network

A semantic segmentation model was developed during this research that processes multi-modal input, incorporating greyscale images and point cloud data from Xenolidar to generate semantic predictions. The primary aspects explored for the semantic segmentation model include the fusion of input modalities and the design of an encoder that efficiently extracts features from road scenes. The preprocessing and semiautomatic annotation methodology developed is detailed in the previous section. This section provides an in-depth description of the developed segmentation network.

The segmentation network architecture is inspired by UNet [9]. To accommodate multi-modal inputs, several modifications were necessary. Various encoder networks were experimented with in this research to determine the most effective models for this use case. The encoders selected for the experimentation were ResNet [7] and EfficientNet [11]. The methods for data fusion and feature extractor design are explained in detail in this section.

Fusion of 2D and 3D Modalities The main aspect addressed is the fusion of the 2D image modality and the 3D point clouds. A late fusion strategy was adopted, where features are extracted from both greyscale images and point clouds from Xenolidar, and then fused in the feature extraction backbone.

Before fusion, the point clouds are converted into an intermediate 2D representation by creating a 2D depth image. This process involves creating an empty array of the same size as the greyscale image and filling it with the depth values from the point cloud. The position of each point in the 2D array is determined by the position of the reflected laser spots on the CMOS sensor, as calibrated by the manufacturer and encoded with the data. Each point in the point cloud includes 2D coordinates, facilitating the conversion to a 2D array. Figure xx shows

the 2D depth image generated from the corresponding 3D point cloud. This 2D depth image is then used in the data fusion with the greyscale image.

The generated 2D depth image is fed into a small network with a few convolutional blocks. The resulting feature vector is concatenated with features from the greyscale images at a later stage in the main feature extraction backbone. The depth feature extractor consists of three convolutional layers followed by average pooling layers, converting a depth image of shape $1\times256\times768$ into $64\times64\times192$.

Proposed Height and Width Attention Block To incorporate the intrinsic features of driving datasets, a custom attention block inspired by HANet [3] was designed and developed. HANet explores the class distribution in the vertical pixel scale, whereas this research proposes a height- and width-driven attention mechanism.

Since the dataset predominantly features highway scenes, a unique class distribution across width and height is present. Vertically, lower sections are expected to have more pixels labelled as roads, while upper sections contain more pixels labelled as sky. Horizontally, the road class is more likely to appear in the middle section, with buildings and vegetation more prominent on the left and right sections.

These observations motivated the design of an additional attention block, named Height- and Width-driven Attention Network (HaWANet). This block is integrated into the later stages of feature extraction and earlier stages of the decoder. HaWANet processes the feature map and, based on height and width-wise contextual information, identifies important features or classes within horizontal and vertical sections, combining this information with features from the main encoder.

The architecture of HaWANet is shown in Figure 4. The input to this attention block is the feature map of the main segmentation network. HaWANet consists of two almost identical parts: the upper section processes height-wise information, while the lower section processes width-wise information. The height attention A_h is calculated in the upper section, and the width attention A_w in the lower section. Then both attention maps are multiplied by the main feature map.

The details of each subsection are as follows:

- **Height-wise pooling (a-h):** Average pooling is applied to the feature map in the width direction to generate a $C_l \times H_l \times 1$ matrix.
- **Downsampling (b-h):** The matrix is downsampled to size $C_l \times H_- \times 1$.
- **Attention map computation (c-h):** Three convolutional layers generate the attention map from the width-wise pooled and downsampled feature map.
- **Upsampling (d-h):** The attention map is upsampled to match the dimensions of the feature map.
- **Positional encoding (e-h):** A sinusoidal positional encoding, as used in HANet, is added to the feature map.

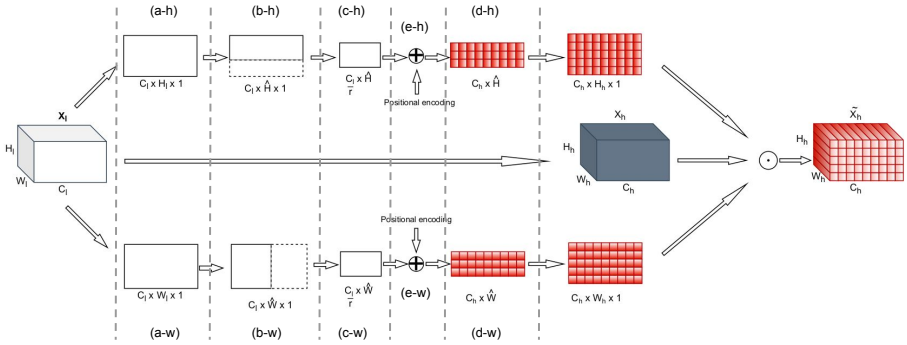

Fig. 4. Height and Width aware Attention Network (HaWANet Architecture)

- **Width-wise pooling (a-w):** Average pooling is applied to the feature map in the height direction to generate a $C_l \times W_l \times 1$ matrix.
- **Downsampling (b-w):** The matrix is downsampled to size $C_l \times W\_ \times 1$.
- **Attention map computation (c-w):** Three convolutional layers generate the attention map from the height-wise pooled and downsampled feature map.
- **Upsampling (d-w):** The attention map is upsampled to match the dimensions of the feature map.
- **Positional encoding (e-w):** A sinusoidal positional encoding is added to the feature map.

HaWANet is an add-on module that can be inserted between different feature extraction blocks in the architecture. The entire encoder-decoder architecture that incorporates data fusion and HaWANet is explained in a subsequent section.

Encoder - Decoder Segmentation Network This research attempts to reuse the concept of UNet [9] to develop a segmentation model. For the encoder part, a ResNet-based feature extractor is used. Performance of multiple variants of ResNet [7] including ResNet-18, ResNet-50, ResNet-101, and EfficientNet [11] was tested. The main modifications made to the encoder part from the standard ResNet are the data fusion scheme and the custom attention block, HaWANet. The model architecture developed is depicted in Figure 5.

Data fusion occurs in the third ResNet block, where features from the greyscale images and the depth image are concatenated. The concatenated feature matrix then progresses through the subsequent stages of feature extraction.

The HaWANet block is added in five positions of the model architecture, specifically in the later stages of feature extraction. This strategy is based on the understanding that the feature matrix is more concise in these stages and can be effectively utilised in the attention layers of HaWANet.

The model follows an encoder-decoder architecture. At intermediate stages, the feature matrix from the encoder section is concatenated with the correspond-

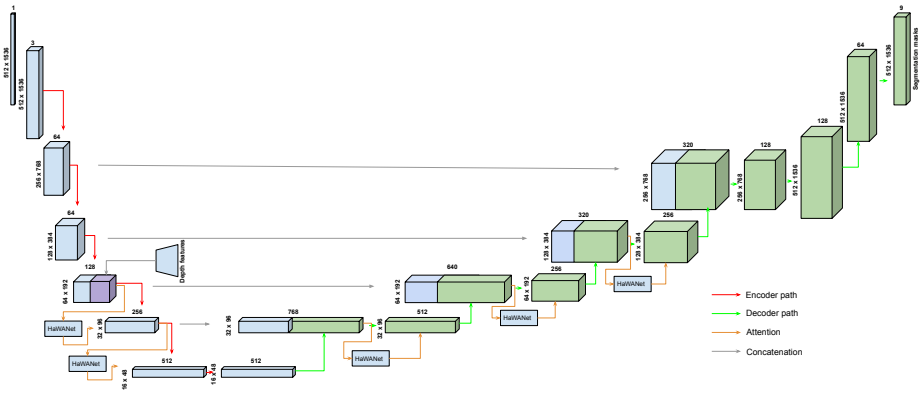

Fig. 5. Encoder - Decoder Network architecture incorporating data fusion and HaWANet attention module

ing sections of the decoder section, inspired by the UNet architecture. Finally, the decoder section outputs the segmentation masks for each class.

The model was trained with approximately 12,000 greyscale images and corresponding depth images for 200 epochs. The loss functions used were a combination of cross-entropy loss and dice loss.

2.3 Model Training

Loss Functions The proposed network architecture was trained using cross entropy [18] and dice loss [10] functions. Cross-entropy loss was implemented to measure the difference between the probability distributions of the predictions and the ground-truth, while the Dice loss was employed to evaluate the segmentation performance.

Data Augmentation This research employed the following data augmentation steps andomly with random intensity to artificially increase the size of the dataset and improve the genaralisation capabilities of the model: rotation from -15 to 15 degrees, horizontal and vertical flip, translation of up to 50 pixels, random signal contrast, and random brightness.

2.4 Design of Experiments

Experiments were designed to study the influence of depth fusion with visual image modality and to test the HaWANet attention module proposed in this research. Multiple models with different backbones were developed to test their performance. Each model was trained with and without depth fusion and tested with the original HANet module and the proposed HaWANet. The backbones

included ResNet-18, ResNet-50, ResNet-101, and EfficientNet. These combinations of backbones, data fusion, and attention modules were grouped and all models were trained.

Evaluated Configurations Different configurations of the proposed architecture were evaluated in this research: baseline UNet, data fusion on UNet, depth fusion HANET block on UNet, HaWANET block on UNet, and finally, depth fusion and HaWANET block on UNet. For feature extraction, four different backbones were evaluated: ResNet18, ResNet50, ResNet101, and EfficientNet.

Evaluation Metrics The ground-truths and the predictions were converted into a bitmap for each class in such a way that a pixel is assigned a value of 1 if it is assigned to that particular class and 0 if it is not, which were then used to calculate *TP, TN, FP* and *FN*. These were then used to computed the *Intersection over Union (IoU)* (also called the *Jaccard Index*).

3 Results

The segmentation model detailed in the previous section was modified with various feature extractors, attention mechanisms, and data fusion schemes, and tested extensively. This section presents the different configurations used to design multiple semantic segmentation models and the training setup. The results of each configuration are described, followed by an overview of the experiments and the final outcomes.

The different configurations were evaluated in terms of accuracy (usig IoU) and speed (using frames per second or FPS), and the scores are presented in Table 1. It was observed that HaWANet with data fusion resulted in the best performance among all configurations evaluated. Furthermore, it was noted that data fusion, through the introduction of the additional modality, improved the performance of HANet, HaWANet, and the baseline model. Table 2 presents the resultant class-wise scores achieved by the best performing model HaWANet with data fusion. Some visual explains of the segmentation results are presented in Figures 6 and 7, for highways or outer road and urban scenes, respectively.

4 Discussion

The primary objective of this research was to develop a multi-modal scene understanding pipeline. The proposed semantic segmentation network utilises an encoder-decoder architecture with a height- and width-driven attention scheme, specifically designed to capture the structural priors of highway road scenes. This architecture integrates depth data from point clouds within the encoder to enhance segmentation accuracy.

The segmentation results presented here demonstrate the model's efficacy in highway scenes. The predicted masks for cars and trucks are highly accurate.

Table 1. Comparison of different configurations with different feature extraction backbones using IoU and FPS

	ResNet 18		ResNet 50		ResNet 101		EfficientNet	
	IoU	FPS	IoU	FPS	IoU	FPS	IoU	FPS
Baseline	0.55±0.04	4.7	0.61±0.07	4.3	0.62±0.04	3.7	0.59±0.05	4.2
Baseline with data fusion	0.57±0.05	4.1	0.63±0.08	3.5	0.64±0.03	3.0	0.61±0.04	3.9
With data fusion and HANet	0.66±0.04	3.8	0.71±0.03	2.9	0.72±0.05	2.5	0.68±0.06	3.7
With HaWANet (without data fusion)	0.65±0.04	3.4	0.70±0.03	2.7	0.71±0.05	2.2	0.70±0.06	3.5
With HaWANet (with data fusion)	0.67±0.04	3.2	0.75±0.03	2.1	**0.78±0.05**	1.9	0.72±0.06	2.9

Table 2. IoU results for individual classes for the best performed model

Class	**IoU**	**Class**	**IoU**	**Class**	**IoU**
Car	0.81	Person	0.70	Sidewalk	0.80
Truck	0.73	Sky	0.75	Building	0.79
Road	0.89	Vegetation	0.78	Bicycle + Motorcycle	0.75

However, the model occasionally misclassifies pixels on the sidewalk as road. The inclusion of point-cloud data, which effectively represents the road's planar nature, is crucial for improving the accuracy of road segmentation.

In urban scenes, despite fewer training samples, the segmentation network performs reasonably well. However, the sidewalk class is sometimes misclassified, likely because the sidewalk and road are at the same level, making them appear similar in the point-cloud data. This misclassification suggests that further refinement is needed in distinguishing these classes.

The limitations of the LiDAR sensor, particularly its maximum range, affect the segmentation of distant objects. Objects far from the sensor lack point cloud data, while the sky often contains invalid data points (e.g., -1 or -9999). Interestingly, these negative values help to segment the sky class effectively, indicating that even seemingly invalid data can provide useful features for segmentation.

The *Person* class has the lowest IoU score among all classes. This poor performance is attributed to the limited number of training samples for this class, primarily because the dataset consists mostly of highway scenes where pedestrians are infrequent. This finding underscores the need for a more balanced dataset or additional data augmentation techniques to improve the segmentation of under-represented classes.

5 Conclusion and Future Work

This research, aimed at addressing the segmentation problem for multimodal sensor data, has presented significant performance improvement over single-modal

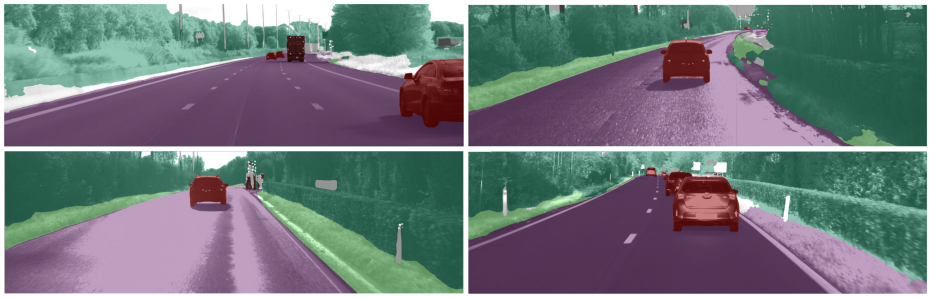

Fig. 6. Predictions of the semantic segmentation model from highways or outer road scenes.

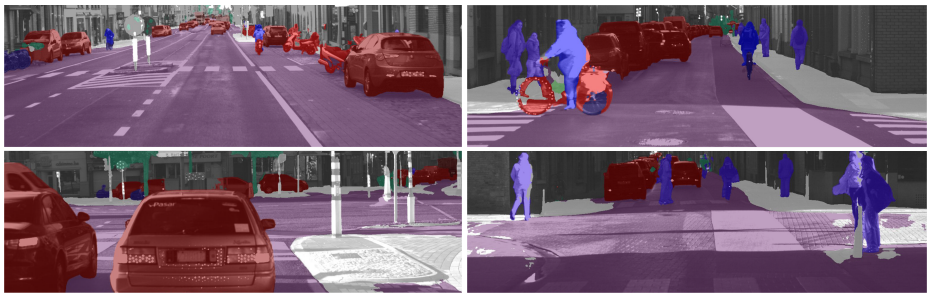

Fig. 7. Predictions of the semantic segmentation model from urban scenes.

approaches. The developed multi-modal segmentation network successfully integrates point cloud data to enhance the understanding of road scenes. The HaWANet attention mechanism effectively captures structural priors, leading to accurate segmentation in various scenarios. However, challenges remain in distinguishing closely related classes and in segmenting under-represented classes.

Future work would focus on expanding the dataset, refining the attention mechanism, and improving the integration of point-cloud data to address these challenges. Additionally, further development and optimisation of object detection models using Xenolidar data can extend the applicability of this research in real-world scenarios.

Acknowledgement

The authors would like to thank Dr Hung Nguyen-Duc, Senior Computer Vision Engineer at Xenomatix NV, Leuven, Belgium, for the support and guidance.

References

1. Chazette, P., Totems, J., Hespel, L., Bailly, J.S.: Principle and Physics of the LiDAR Measurement, pp. 201–247 (12 2016). https://doi.org/10.1016/

B978-1-78548-102-4.50005-3

2. Chen, B.S., Foster, P., Warkentine, R.: Research and development of VCSEL-based optical sensors in industrial applications. In: Choquette, K.D., Lei, C. (eds.) Vertical-Cavity Surface-Emitting Lasers V. vol. 4286, pp. 210 – 218. International Society for Optics and Photonics, SPIE (2001). https://doi.org/10.1117/12.424806, https://doi.org/10.1117/12.424806
3. Choi, S., Kim, J.T., Choo, J.: Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9373–9383 (2020)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995). https://doi.org/https://doi.org/10.1007/BF00994018
5. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013)
6. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13. pp. 213–228. Springer (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Li, Y., Ibanez-Guzman, J.: Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. IEEE Signal Processing Magazine **37**(4), 50–61 (2020). https://doi.org/10.1109/MSP.2020.2973615
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
10. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. pp. 240–248. Springer (2017)
11. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
12. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. International Journal of Computer Vision **128**(5), 1239–1285 (2020)
13. Valada, A., Oliveira, G.L., Brox, T., Burgard, W.: Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In: 2016 international symposium on experimental robotics. pp. 465–477. Springer (2017)
14. Wada, K.: Labelme: Image Polygonal Annotation with Python. https://doi.org/10.5281/zenodo.5711226, https://github.com/wkentaro/labelme
15. Wang, D., Watkins, C., Xie, H.: Mems mirrors for lidar: A review. Micromachines **11**(5) (2020). https://doi.org/10.3390/mi11050456, https://www.mdpi.com/2072-666X/11/5/456

16. Warren, M., Block, M., Dacha, P., Carsonn, R., Podva, D., Helms, C., Maynard, J.: Low-divergence high-power vcsel arrays for lidar application. p. 14 (02 2018). https://doi.org/10.1117/12.2290937
17. Xenomatix: Xenomatix solidstate lidar scanner (2022), https://xenomatix.com/solid-state-lidar/, [Online; accessed 19-April-2022]
18. Yi-de, M., Qing, L., Zhi-bai, Q.: Automated image segmentation using improved pcnn model based on cross-entropy. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004. pp. 743–746 (2004). https://doi.org/10.1109/ISIMP.2004.1434171
19. Zhang, Y., Sidibé, D., Morel, O., Mériaudeau, F.: Deep multimodal fusion for semantic image segmentation: A survey. Image and Vision Computing **105**, 104042 (2021). https://doi.org/https://doi.org/10.1016/j.imavis.2020.104042, https://www.sciencedirect.com/science/article/pii/S0262885620301748