Adaptive-Mask Fusion Network for Segmentation of Drivable Road and Negative Obstacle With Untrustworthy Features

Zhen Feng[1,2], Yuchao Feng[1], Yanning Guo[2], and Yuxiang Sun[1,*]

2023 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3503-4691-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IV55152.2023.10186628

***Abstract*— Segmentation of drivable roads and negative obstacles is critical to the safe driving of autonomous vehicles. Currently, many multi-modal fusion methods have been proposed to improve segmentation accuracy, such as fusing RGB and depth images. However, we find that when fusing two modals of data with untrustworthy features, the performance of multi-modal networks could be degraded, even lower than those using a single modality. In this paper, the untrustworthy features refer to those extracted from regions (e.g., far objects that are beyond the depth measurement range) with invalid depth data (i.e., 0 pixel value) in depth images. The untrustworthy features can confuse the segmentation results, and hence lead to inferior results. To provide a solution to this issue, we propose the adaptive-mask fusion Network (AMFNet) by introducing adaptive-weight masks in the fusion module to fuse features from RGB and depth images with inconsistency. In addition, we release a large-scale RGB-depth dataset with manually-labeled ground truth based on the NPO dataset for drivable roads and negative obstacles segmentation. Extensive experimental results demonstrate that our network achieves state-of-the-art performance compared with other networks. Our code and dataset are available at: https://github.com/lab-sun/AMFNet.**

I. Introduction

Segmentation of drivable roads and negative obstacles is a fundamental capability for autonomous vehicles. Although vehicles can generally pass small negative obstacles on roads, negative obstacles are still potential threats to vehicles. Especially when vehicle speed is fast or negative obstacles are large, severe accidents, such as roll over, could happen [1]. Accurate segmentation results of drivable roads and negative obstacles could serve as input data for downstream tasks, such as path planning [2], to avoid potential accidents.

Many single-modal (e.g., using only RGB images) networks have been proposed for the segmentation of drivable roads and negative obstacles [3], [4]. To improve the segmentation performance, multi-modal networks based on RGB-depth (RGB-D) fusion [5]–[7] and RGB-disparity fusion [6], [8] have been proposed. Although these networks have achieved acceptable results, we find that when there are a large number of pixels in depth images without valid depth information (i.e., pixel value 0 in depth images), the segmentation performance cannot been improved or even inferior to the performance with a single RGB modality. We call the regions with the pixel value 0 in depth images as untrusted regions. The value 0 in depth images indicates that the depth information of the object cannot be measured (e.g., out of the depth measurement range), rather than indicating that the distance between the object and the camera is 0. The features extracted from untrusted regions could not represent the real features of the environment, so we call the features as untrustworthy features. The untrustworthy features could confuse the segmentation results since there are valid features from the other modality.

To provide a solution to this issue, we propose a novel adaptive-mask fusion network (AMFNet) with adaptive-mask fusion (AMF) modules. To this end, we generate mask images from depth images to distinguish trusted and untrusted regions. The AMF module is used to generate adaptive-weight masks for RGB and depth feature maps to reduce the influence caused by untrustworthy features during fusion. We also release a large-scale RGB-depth (RGB-D) dataset with manually-labeled ground truth for drivable roads and negative obstacles segmentation. Our contributions are summarized as follows:

- We propose an adaptive-mask fusion (AMF) module to reduce the influence of untrustworthy features during feature fusion.
- We proposed a novel fusion network named AMFNet with the AMF modules for the segmentation of drivable roads and negative obstacles.
- We release a large-scale RGB-D dataset based on the NPO dataset[1]. Our dataset consists of 8,752 RGB-D images with manually-labeled ground truth for the segmentation of drivable roads and negative obstacles.

II. Related Works

A. Semantic Segmentation Networks

Chen *et al.* [9] designed DeepLabV3+ with atrous convolution in encoder-decoder structure for semantic segmentation. Azad *et al.* [10] introduced attention modules into DeepLabV3+ to propose Att-Deeplabv3+. Recently, many Transformer-based semantic segmentation networks have been proposed. Hatamizadeh *et al.* [11] combined the U-shaped structure and Transformer structure to design Swin UNETR for medical image segmentation. Yuan *et al.* [12] proposed CTC-Net for medical image segmentation with a convolutional neural networks-based encoder and a transformer-based encoder.

To improve semantic segmentation accuracy, many multi-modal fusion networks have been proposed. Hazirbas *et al.*

[1]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (email: zfeng94@outlook.com; yuchao.feng@connect.polyu.hk; yx.sun@polyu.edu.hk, sun.yuxiang@outlook.com).

[2]Harbin Institute of Technology, Harbin, Heilongjiang, China (email: zfeng94@outlook.com; guoyn@hit.edu.cn).

*Corresponding author: Yuxiang Sun.

[1]https://github.com/lab-sun/InconSeg

[13] proposed FuseNet to fuse RGB images and depth images for semantic segmentation. Sun *et al.* [14] proposed RTFNet to fuse RGB images and thermal images for the segmentation of urban scenes. Zhou *et al.* [15] proposed FRNet to fuse RGB images and depth images with a cross-level enriching module in the encoder.

B. Semantic Segmentation of Drivable Road

Wang *et al.* [16] proposed a normal inference module (NIM) for the depth image to improve the performance of drivable areas and road anomaly detection. The performance of several networks embedded with NIM has been improved. Fan *et al.* [7] proposed a surface normal estimator for depth images and designed RoadSeg to fuse the output of the surface normal estimator and RGB images. Kothandaraman *et al.* [17] proposed an unsupervised method to segment roads under adverse weather conditions. Chin *et al.* [18] proposed transformer-based OFF-Net to fuse LiDAR point cloud and RGB image. They also released the ORFD dataset with 12, 198 pairs of LiDAR point cloud and RGB images.

C. Semantic Segmentation of Negative Obstacles

Fan *et al.* [8] proposed AA-RTFNet by combining RTFNet and attention modules to fuse RGB images and disparity images. They also released the Pothole-600 dataset for the segmentation of potholes. Feng *et al.* [5] proposed MAFNet to fuse RGB images and disparity images for the segmentation of potholes. Masihullah *et al.* [19] combined attention modules with DeepLabV3+ to segment roads and potholes.

Although the aforementioned multi-modal fusion networks have achieved acceptable results, they all ignore the influence of untrustworthy features. We find that untrustworthy features could degrade the fusion performance. So, in this work, we propose the AMF module to reduce the influence of untrustworthy features.

III. The Proposed Method

A. The Overall Architecture

Fig. II shows the overall architecture of our proposed AMFNet. Our AMFNet is designed with the structure. It consists of a five-stage RGB encoder, a five-stage depth encoder, and a five-stage decoder. There are also 5 AMF modules in the RGB encoder. Each AMF module is placed behind each stage of the RGB encoder. The RGB encoder and depth encoder are borrowed from BotNet-50 [20]. Depth images are used to generate masks with a threshold of 0. When the value of a pixel in the depth image is greater than 0, the value of the pixel in the mask is 1. We believe that the value 0 in the depth image is untrustworthy because the distance between the real environment point and the camera is not 0. The mask is a map used to distinguish trustworthy pixels from untrustworthy pixels. The mask is downsampled to generate five different masks (i.e., M_1, M_2, M_3, M_4, and M_5), which have the same resolution as the outputs of the 5 stages of the RGB encoder. The M_n mask is fed into the n-th AMF module, where $n \in [0, 5]$. The outputs of the same level stages of the RGB and depth encoders are fed into the same level AMF module. The output of the n-th AMF module is fed into the $(n + 1)$-th stage of the RGB encoder and fused into the output of the $(n + 1)$-th stage of the decoder by element-wise addition.

B. The AMF Module

The structure of the AMF module is shown in Fig. 2. The n-th AMF module has three inputs: the output of the n-th stage of the RGB encoder (RGB feature map), the output of the n-th stage of the depth encoder (depth feature map), and the n-th mask M_n. In each AMF module, the mask is first fed into an adaptive mask generation (AMG) module to generate two adaptive-weight masks for the RGB feature map and depth feature map. Secondly, the adaptive-weight masks are fused with the RGB feature map and depth feature map by element-wise multiplication, and then fused to generate the result of the fusion of the RGB feature map and depth feature map by element-wise addition. Finally, the weights of the fusion result of the RGB feature map and depth feature map are adjusted by a channel attention block and a spatial attention block. In the channel attention block, the fusion result is passed through an adaptive average pooling layer, a fully connected (FC)-batch normalization (BN)-ReLU layer, an FC layer, and a Sigmoid layer sequentially to generate the weights of each channel. The weights of each channel are fused into the result of the fusion to generate an adjusted result by element-wise multiplication. In the spatial attention block, the adjusted result is passed through a convolutional layer and a Sigmoid layer sequentially to generate spatial weights. The spatial weights are fused into the adjusted result to generate the output of AMF module by element-wise multiplication.

The main purpose of the AMF module is to divide the feature map into trusted regions and untrusted regions according to the mask. In the trusted regions, RGB features and trustworthy depth features are fused by adaptive weights. In the untrusted regions, the untrustworthy features of the depth images are discarded, and the RGB features are directly used as the fusion result. The AMG module is designed to achieve this purpose. The AMG module has three inputs: the mask, the RGB feature map, and the depth feature map. The three inputs have the same resolution. Firstly, the RGB feature map and depth feature map are concatenated together. Secondly, the concatenated feature map is passed through an adaptive average pooling layer, two FC-BN-ReLU layers, and an FC-BN-Softmax layer. The outputs of the FC-BN-Softmax layer are two weights for the RGB feature map and depth feature map. Thirdly, the weight for the depth feature map is fused with the mask by element-wise multiplication to generate the mask for the depth feature map. Finally, the depth-feature-map mask is subtracted from an all-one matrix to generate the RGB-feature-map mask.

C. The Decoder

The decoder consists of five stages with the same structure. Fig. II shows the structure of one stage. The input of one stage is first fed into a dual residual block. Secondly, the

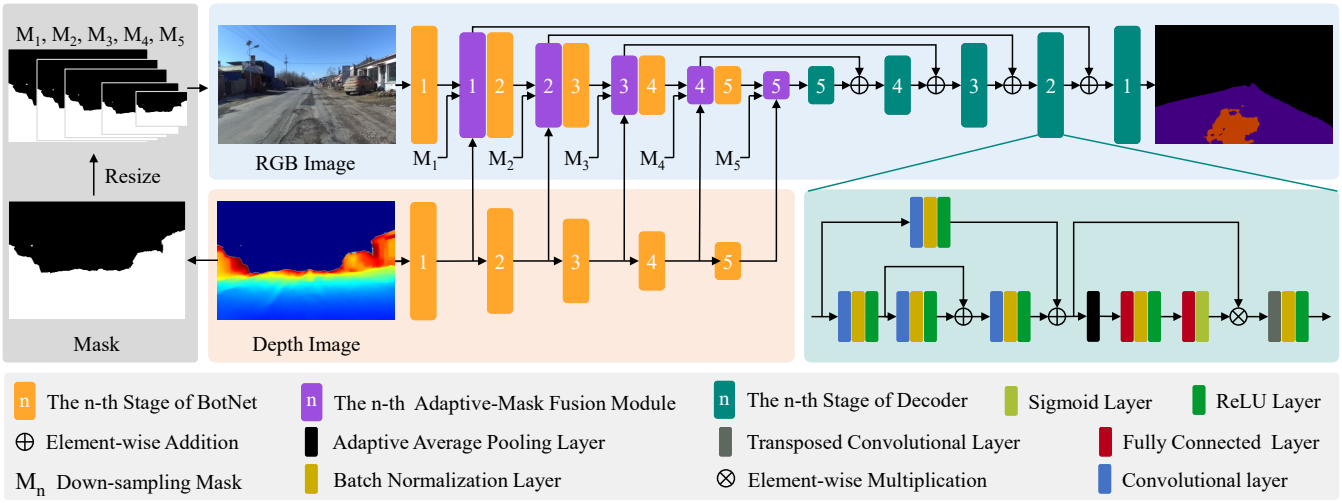

Fig. 1: The overall architecture of our proposed AMFNet. Our AMFNet adopts the two-encoder-one-decoder structure: a five-stage RGB encoder, a five-stage depth encoder, and a five-stage decoder. The encoder is adopted from BotNet-50 [20]. Our proposed adaptive-mask fusion (AMF) modules are placed behind each stage of the RGB encoder. The mask is generated by thresholding the depth image with the pixel value 0. Five different masks (i.e., M_1, M_2, M_3, M_4, and M_5) with the same resolution as the outputs of the 5 stages of the RGB encoder are generated by downsampling with the nearest neighbor method. The figure is best viewed in color.

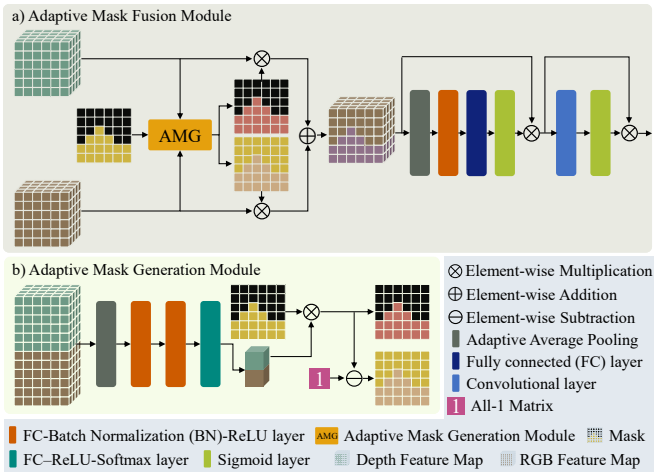

Fig. 2: The structure of an AMF module. Both AMF and adaptive mask generation (AMG) Module have the same three inputs: an RGB feature map, a depth feature map, and a mask. The outputs of the AMG are two adaptive-weight masks for the RGB feature map and depth feature map.

output of the dual residual block is fed into a channel attention block to adjust the weights of each channel. Finally, a transposed Convolution-BN-ReLU (CBR) layer is used to generate the output of the stage.

There are four CBR layers in the dual residual block. The input is fed into the first CBR layer and the fourth CBR layer. The output of the first CBR layer is fed into the second CBR layer and fused with the output of the second CBR layer. The fusion result is fed into the third CBR layer. The outputs of the third and the fourth CBR layer are fused together as the output of the dual residual block.

IV. Experimental Results and Discussions

A. The Dataset

As aforementioned, large-scale multi-modal datasets with drivable roads and negative obstacles are very limited. Therefore, we release a large-scale RGB-D dataset based on the NPO dataset for the segmentation of drivable roads and negative obstacles. The raw images of the NPO dataset were recorded with a ZED stereo camera mounted on a vehicle. There are 20 image sequences in the raw data of the NPO dataset. We manually label one image per 5 images in some image sequences that include nearly 44,000 image pairs (left images, right images, and depth images) with $1,242 \times 2,208$ resolution. So, in total, 8,752 images are labelled in our dataset. To alleviate the annotation task, we directly use the masks of negative obstacles in the NPO dataset as the masks for our annotation. We name our dataset as Drivable Roads and Negative Obstacles (DRNO) dataset. There are various lighting conditions, weather conditions, and scenes in our dataset, such as normal lighting, large areas of shadow, snowy, sunny, cloudy, urban scenes, and rural scenes. There are also various road surface types in the data set, such as water, snow, and normal road surfaces.

To the best of our knowledge, our DRNO dataset is the largest dataset for semantic segmentation of drivable roads and negative obstacles. Some samples of our dataset are shown in Fig. 3. In our DRNO dataset, 8,752 images include drivable roads, and 748 images include negative obstacles.

B. Training Details

Our AMFNet is implemented with PyTorch. The network is trained and tested on a PC with NVIDIA RTX 3090 (24 GB RAM) graphics card. The parameters of the first four encoder stages of AMFNet are initialized with the pre-trained

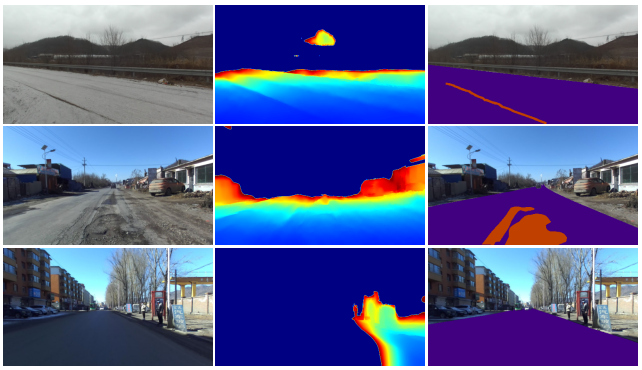

Fig. 3: Sample images in our DRNO dataset. We visualize the depth images with the *jet* color map. Depth values increase from red to blue. ■ and ■ represent drivable roads and negative obstacles, respectively. The figure is best viewed in color.

weight from PyTorch. We employ the stochastic gradient descent (SGD) optimizer to train the network. The initial learning rate is set to 0.01, the momentum is set to 0.9, and the decay strategy is set to 0.95.

We randomly split our dataset into three sets: training (4,376 pairs of RGB-D images), validation (2,188 pairs of RGB-D images), and testing (2,188 pairs of RGB-D images). To trade-off training speed and network performance, we reduce the resolution of the images to 288×512 during training and testing. The unlabelled background is treated as a class during training and testing.

C. Ablation Study

We conduct ablation study to check the benefits of the AMF module and choose the optimal structure for our AMFNet. Firstly, we place one AMF module behind the different stages of the RGB encoder to design variants. For example, an AMF module is placed behind the 5-th stage of the RGB encoder to fuse RGB and depth feature maps in a variant. Secondly, we design variants by placing AMF modules behind different stages of the RGB encoder. For example, AMF modules are placed behind the last two stages of the RGB encoder in a variant. In all the variants, the outputs of the same-level stages without AMF modules are fused by element-wise addition. The metrics Mean Accuracy (mAcc), mean F-score (mF1), and mean Intersection-over-Union (mIoU) [5] are used to evaluate the performance of all the variants.

The results are displayed in Tab. I. From the results, we can find that the variant without any AMF module achieves inferior results. Comparing variant A to variant F, we can find that no matter where AMF is placed, it can always improve the performance of the network. Comparing the four variants G, H, I, and J, more AMF modules in one variant lead to better performance. This shows that our proposed AMF module can remove untrustworthy features in the fusion process, thus improving the fusion performance. Based on the experimental results, we place five AMF modules behind each stage of the RGB encoder in our AMFNet.

TABLE I: The results (%) of the ablation study. '✓' means AMF module is placed behind the n-th stage of the RGB encoder. '–' means that the outputs of the n-th stage of the RGB encoder and depth encoder are fused with element-wise addition. The best results are highlighted in bold font.

No.	Stage					mAcc	mIoU	mF1
	1st	2nd	3rd	4th	5th			
(A)	–	–	–	–	–	67.57	66.54	69.18
(B)	–	–	–	–	✓	69.96	67.00	69.18
(C)	–	–	–	✓	–	68.36	67.05	69.87
(D)	–	–	✓	–	–	68.48	66.86	69.72
(E)	–	✓	–	–	–	67.79	66.58	69.22
(F)	✓	–	–	–	–	67.99	66.76	69.49
(G)	–	–	–	✓	✓	68.89	67.21	70.15
(H)	–	–	✓	✓	✓	69.10	67.27	70.24
(I)	–	✓	✓	✓	✓	69.98	67.99	71.34
(J)	✓	✓	✓	✓	✓	**70.60**	**68.39**	**71.99**

D. Comparative Study

We compare our proposed AMFNet with the well-known networks: FuseNet [13], RTFNet [14], AA-RTFNet [8], RoadSeg [7], SS-SFDA [17], MAFNet [5], OFF-Net [18], and FRNet [15]. We use our DRNO dataset to train and test the networks. To illustrate the impact of untrustworthy features on existing multi-modal fusion networks, we also train and test the aforementioned multi-modal networks without RGB encoders or depth encoders. We also removed the fusion module from the multi-modal network during training and testing. In other words, each multi-modal fusion network is trained and tested by three different modalities, namely single RGB modality, single depth modality, and RGB-depth fusion modality. The mAcc, mF1, and mIoU are also used as metrics to evaluate the performance of our AMFNet and these networks. In addition, the Acc, F1, and IoU of each class (i.e., background, drivable road, and negative obstacles) are also used as evaluation metrics.

1) Quantitative Results: The results of all networks are displayed in Tab. II. Comparing the results of each multi-modal fusion network, we can find that the results of the single RGB modality of all networks are better than the results of the RGB-depth fusion modality, except our network. These networks fuse untrustworthy features as general features, thus degrading the results of multi-modal fusion. Our proposed network reduces the influences of the untrustworthy features through the AMF module, making the results of the RGB-depth fusion modality better than that of the single-RGB modality. These results confirm our conjecture that untrustworthy features hinder multi-modal fusion. Comparing all the results, our network almost achieves the best results in terms of all metrics. This illustrates the superiority of our AMFNet.

2) Qualitative Demonstrations: Some sample qualitative results of the top-3 multi-modal fusion networks (i.e., our AMFNet, MAFNet, and RoadSeg) with the best mIoU metric are shown in Fig. 4. From the third column of the results, we can see that the snow cover on the road confuses the segmentation of negative obstacles. MAFNet and RoadSeg

TABLE II: The comparative results (%) on the testing set of our DRNO dataset. 'Modality' means the type of modality for training networks. 'Year' means the published year of networks. 'RGB & Depth' means the network is trained and tested with RGB-depth fusion modality. The results demonstrate the superiority of our AMFNet. The best results are highlighted in bold font.

Network	Year	Modality	Background			Drivable Road			Negative Obstacles			mAcc	mIoU	mF1
			Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1			
FuseNet [13]	2016	Depth	98.50	97.40	98.68	97.67	94.61	97.23	0.00	0.00	0.00	65.39	64.00	65.30
		RGB	98.76	97.67	98.82	97.70	95.15	97.51	0.00	0.00	0.00	65.49	64.27	65.45
		RGB & Depth	98.95	97.59	98.78	97.13	94.96	97.42	0.00	0.00	0.00	65.36	64.18	65.40
RTFNet [14]	2019	Depth	99.20	97.65	98.81	96.71	95.02	97.45	1.39	1.28	2.52	65.77	64.65	66.26
		RGB	99.48	97.92	98.95	96.69	95.56	97.73	7.82	6.25	11.77	68.00	66.58	69.48
		RGB & Depth	99.50	97.90	98.94	96.56	95.47	97.68	7.07	4.21	8.08	67.71	65.86	68.23
AA-RTFNet [8]	2020	Depth	99.11	97.62	98.80	96.84	94.98	97.43	1.25	1.07	2.11	65.73	64.56	66.11
		RGB	99.47	98.13	99.05	97.06	95.93	97.92	14.14	7.11	13.27	70.22	67.06	70.08
		RGB & Depth	99.45	98.06	99.02	97.01	95.81	97.86	7.11	4.37	8.38	67.86	66.08	68.42
RoadSeg [7]	2020	Depth	98.30	96.44	98.19	96.02	92.66	96.19	0.00	0.00	0.00	64.78	63.03	64.79
		RGB	98.92	98.19	99.09	**98.44**	96.18	98.06	8.39	6.83	12.78	68.58	67.07	69.97
		RGB & Depth	99.28	98.09	99.04	97.44	95.91	97.91	7.34	4.84	9.24	68.02	66.28	68.73
SS-SFDA [17]	2021	Depth	98.47	96.62	98.28	96.05	93.00	96.37	0.00	0.00	0.00	64.84	63.21	64.88
		RGB	98.59	97.07	98.51	96.76	93.91	96.86	0.84	0.79	1.57	65.40	63.92	65.65
		RGB & Depth	98.61	96.89	98.42	96.34	93.54	96.66	0.07	0.07	0.14	65.01	63.50	65.08
MAFNet [5]	2022	Depth	98.57	96.95	98.45	96.55	93.67	96.73	0.00	0.00	0.00	65.04	63.54	65.06
		RGB	99.44	98.14	99.06	97.26	96.04	97.98	8.11	7.35	13.69	68.27	67.17	70.24
		RGB & Depth	99.51	98.14	99.06	97.09	96.01	97.97	4.90	4.03	7.76	67.17	66.06	68.26
OFF-Net [18]	2022	Depth	96.71	91.95	95.81	89.34	83.57	91.05	0.00	0.00	0.00	62.02	58.51	62.29
		RGB	98.17	96.60	98.27	96.66	93.02	96.38	0.00	0.00	0.00	64.94	63.21	64.88
		RGB & Depth	97.51	96.37	98.15	97.56	92.68	96.20	0.00	0.00	0.00	65.02	63.02	64.78
FRNet [15]	2022	Depth	99.10	97.76	98.87	97.18	95.29	97.59	0.00	0.00	0.00	65.43	64.35	65.49
		RGB	99.42	97.92	98.95	96.83	95.58	97.74	6.64	5.89	11.13	67.63	66.47	69.27
		RGB & Depth	**99.52**	98.02	99.00	96.78	95.72	97.81	7.46	4.62	8.83	67.92	66.12	68.55
AMFNet (Ours)		Depth	99.07	97.58	98.78	96.85	94.90	97.39	1.08	0.99	1.97	65.67	64.49	66.04
		RGB	99.26	98.25	99.12	97.88	96.30	98.12	8.86	8.01	14.82	68.67	67.52	70.69
		RGB & Depth	99.25	**98.40**	**99.19**	98.17	**96.57**	**98.26**	**14.39**	**10.20**	**18.51**	**70.60**	**68.39**	**71.99**

incorrectly segment negative obstacles due to the influence of the snow cover. However, our AMFNet correctly segments the drivable road. From the results in the seventh column, we can see that the water and shadows on the road seriously degrade the segmentation performance of the drivable road. Our AMFNet also achieves the best results among the three networks. From the results in the fifth column, we can see that our AMFNet correctly segments most areas of negative obstacles. The results illustrate the superiority of our AMFNet.

V. Conclusions and Future Work

We proposed here a novel network AMFNet with the AMF module for the segmentation of drivable roads and negative obstacles. Our proposed network addresses the degradation of fusion performance caused by untrustworthy features extracted from depth images. We generated masks from depth images as the input of AMF module. The AMF module generates two adaptive-weight masks for the RGB feature map and depth feature map to reduce the influence of untrustworthy features. In addition, we released a large-scale RGB-D dataset with pixel-level ground truth of drivable roads and negative obstacles for this task. The experimental results show that our proposed AMFNet achieves better performance than single-RGB modality in the presence of untrustworthy features during fusion. However, there are also some limitations in our AMFNet. For example, the segmentation accuracy of the class *negative obstacles* is low. We would like to improve the weight of the loss of the *negative obstacles* during training in the future.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62003286, in part by Zhejiang Lab under Grant 2021NL0AB01, and in part by the HK PolyU Start-up Fund under Grant P0034801.

References

[1] T. Verster and E. Fourie, "The good, the bad and the ugly of south african fatal road accidents," *South Afr. J. Sci.*, vol. 114, no. 7/8, pp. 63–69, Jul. 2018.

[2] G. Cheng and J. Y. Zheng, "Sequential semantic segmentation of road profiles for path and speed planning," *IEEE Trans. on Intell. Transp. Sys.*, vol. 23, no. 12, pp. 23 869–23 882, 2022.

[3] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 607–12 616.

[4] V. Pereira, S. Tamura, S. Hayamizu, and H. Fukai, "Semantic segmentation of paved road and pothole image using u-net architecture," in *Proc. Int. Conf. Adv. Inf., Concepts Theory Appl.*, 2019, pp. 1–4.

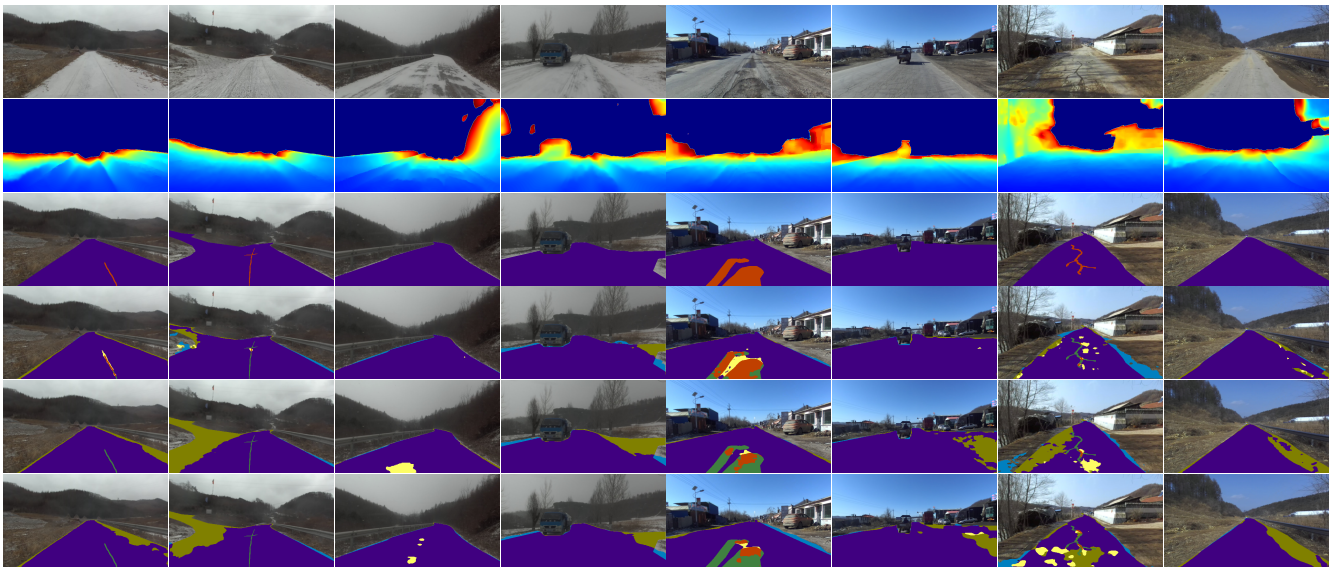

Fig. 4: Sample qualitative demonstrations of the top-3 multi-modal fusion networks with the best mIoU metric. The 4-th row to the 6-th row are respectively demonstrations of our AMFNet, MAFNet [5], and RoadSeg [7]. ■, ■, and ■ represent negative obstacles, the false positive of negative obstacles, and the false negative of negative obstacles. ■, ■, and ■ represent drivable roads, the false positive of drivable roads, and the false negative of drivable roads, respectively. The figure is best viewed in color.

[5] Z. Feng, Y. Guo, Q. Liang, M. U. M. Bhutta, H. Wang, M. Liu, and Y. Sun, "Mafnet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles," *IEEE Trans. on Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[6] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10 750–10 760, 2021.

[7] R. Fan, H. Wang, P. Cai, and M. Liu, "Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[8] R. Fan, H. Wang, M. J. Bocus, and M. Liu, "We learn better road pothole detection: from attention aggregation to adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 285–300.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[10] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 251–266.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*, 2022, pp. 272–284.

[12] F. Yuan, Z. Zhang, and Z. Fang, "An effective cnn and transformer complementary network for medical image segmentation," *Pattern Recognition*, vol. 136, p. 109228, 2023.

[13] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

[14] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, 2019.

[15] W. Zhou, E. Yang, J. Lei, and L. Yu, "Frnet: Feature reconstruction network for rgb-d indoor scene parsing," *IEEE J. Sel. Topics Signal Process*, vol. 16, no. 4, pp. 677–687, 2022.

[16] H. Wang, R. Fan, Y. Sun, and M. Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst.*, 2020, pp. 2706–2711.

[17] D. Kothandaraman, R. Chandra, and D. Manocha, "Ss-sfda: Self-supervised source-free domain adaptation for road segmentation in hazardous environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3049–3059.

[18] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai, "Orfd: A dataset and benchmark for off-road freespace detection," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 2532–2538.

[19] S. Masihullah, R. Garg, P. Mukherjee, and A. Ray, "Attention based coupled framework for road and pothole segmentation," in *Proc. Int. Conf. Pattern Recognit)*, 2021, pp. 5812–5819.

[20] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16 519–16 529.