HOKKAIDO UNIVERSITY

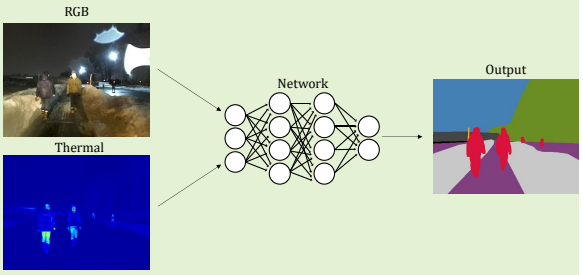
Title	Multi-Modal Sensor Fusion-Based Semantic Segmentation for Snow Driving Scenarios
Author(s)	Vachmanus, Sirawich; Ravankar, Ankit A.; Emaru, Takanori et al.
Citation	IEEE sensors journal, 21(15), 16839-16851 https://doi.org/10.1109/JSEN.2021.3077029
Issue Date	2021-08-01
Doc URL	https://hdl.handle.net/2115/82600
Rights	© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Type	journal article
File Information	Final-Manuscript.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Multi-modal Sensor Fusion-based Semantic Segmentation For Snow Driving Scenarios

Sirawich Vachmanus, *Student Member, IEEE*; Ankit A. Ravankar, *Member, IEEE*; Takanori Emaru, *Member, IEEE*; and Yukinori Kobayashi

***Abstract*— In recent years, autonomous vehicle driving technology and advanced driver assistance systems have played a key role in improving road safety. However, weather conditions such as snow pose severe challenges for autonomous driving and are an active research area. Thanks to their superior reliability, the resilience of detection, and improved accuracy, advances in computation and sensor technology have paved the way for deep learning and neural network–based techniques that can replace the classical approaches. In this research, we investigate the semantic segmentation of roads in snowy environments. We propose a multi-modal fused RGB-T semantic segmentation utilizing a color (RGB) image and thermal map (T) as inputs for the network. This paper introduces a novel fusion module that combines the feature map from both inputs. We evaluate the proposed model on a new snow dataset that we collected and on other publicly available datasets. The segmentation results show that the proposed fused RGB-T input can segregate human subjects in snowy environments better than an RGB-only input. The fusion module plays a vital role in improving the efficiency of multiple input neural networks for person detection. Our results show that the proposed network can generate a higher success rate than other state-of-the-art networks. The combination of our fused module and pyramid supervision path generated the best results in both mean accuracy and mean intersection over union in every dataset.**



***Index Terms*— machine learning, semantic segmentation, thermal camera, data fusion**

I. Introduction

INCLEMENT weather conditions such as fog, rain, smoke, or snow can severely hamper drivers' visibility and pose a serious risk of accidents. In the United States, casualties due to vehicular crashes total more than 1.5 million annually, with 800,000 injuries [1]. Weather conditions can significantly change road surface dynamics, causing delays and warnings [2]. The identical accident risk based on the road type from 2014–2016 in Finland was the highest on the slushy, snowy roads and higher on expressways than two-lane and multi-lane roads [3]. Similarly, in Japan, wet and frozen road surfaces, even under clear weather conditions, can cause severe traffic accidents. Moreover, snowy weather conditions also have the highest number of injuries compared with other conditions [4]. Fig.1 shows some examples of the road environment in the Hokkaido prefecture of Japan. This region receives the highest annual snowfall and is covered with snow for up to four months, meaning there is a high probability of accidents and vehicle slippage. Such situations cause severe delays in transportation and loss of business. Moreover, driving in such challenging conditions with poor visibility and slippery roads is very stressful for drivers.

During the winter season, road-heating services are generally not available because of their higher costs, and most city governments employ snow graders to sweep snow off the surface, hence making piles of snow on the roadside. The snow piles are cleared out by snow removal machines during the night. This process is hazardous for pedestrians and other vehicles in the surrounding areas. The bottom-right image in Fig.1 shows snowy roads at night, indicating that it is challenging to distinguish pedestrians because of poor visibility.

Recently, autonomous vehicle technology has shown a great deal of promise in improving road safety. Under clear weather conditions, autonomous driving systems can navigate the

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP20K04392, entitled "Robust SLAM by Non-Uniform UGV/UAV Groups for Large Field Management." Part of this work was carried out by a joint research with Sapporo City, "Alternative Technology for Traffic Guides in Sidewalk Snow Removal." In addition, we received a great deal of cooperation from the Snow Countermeasures Office of the Sapporo City Construction Bureau in carrying out the experiment.

S. Vachmanus is a PhD student at the research laboratory of Robotics and Dynamics of Hokkaido University, Hokkaido, Japan (e-mail: vachmas@gmail.com).

A. A. Ravankar is currently a Research Associate/ Lecturer at the Department of Robotics, Division of Mechanical Engineering, Tohoku University, Japan. (e-mail: ankit@srd.mech.tohoku.ac.jp).

T. Emaru is currently an Associate Professor of the Department of Human Mechanical System and Design Engineering, Research laboratory of Robotics and Dynamics of Hokkaido University, Hokkaido, Japan. (e-mail: emaru@eng.hokudai.ac.jp).

Y. Kobayashi is currently the President of the National Institute of Technology, Tomakomai College, Hokkaido, Japan (e-mail: kobay@eng.hokudai.ac.jp).

Fig. 1. Snowy environment in Hokkaido, Japan.

vehicle with high accuracy. In contrast to indoor environments, outdoor environments are far more challenging because of the lack of features that are supported working under such conditions [5], [6]. Moreover, challenging weather conditions can cause poor visibility, directly affecting the accuracy of the vehicle's perception, which is one of the three main functions of an autonomous system. When it comes to perception, many sensors, such as cameras, LiDAR (light detection and ranging), and thermal imaging, have been used to detect the environment [7], [8], [9]. The research on road perception and recognition has gained significant importance in recent years, and many methods have been developed to support drivers' perception and recognition [10], [11]. The classical methods in road detection are based on image processing techniques [12], [13], [14]. For road recognition, image processing has become a preferred method because of the low cost of sensors, higher performance, and better computation.

At the moment, deep learning and neural network–based methods have replaced the classical approaches because of the reliability and resilience of detection. Deep learning techniques have been developed in many fields, such as fingerspelling identification [15], [16] or traffic recognition [17]. Many of the new studies employing machine learning and deep learning methods have shown improved accuracy in detection compared with the classical approaches. Semantic segmentation is one learning technique gaining a lot of attention when it comes to understanding objects in the image at the pixel level [18]. By employing millions of images for training, this method can detect the road and other objects in the images simultaneously, including objects such as traffic signs, electric poles, or even pedestrians. However, the most common methods that work well under normal conditions will fail in snowy environments. The challenge is to detect environmental features under the snow cover, an issue that arises because of the minimum separation between color pixels. In addition, snow and rain can obscure sensors, hide road signs and lane markings, and affect the car's performance. Bad weather such as snow represents a difficult test for artificial intelligence algorithms. For example, it is almost impossible to separate the road and sidewalk when both are under snow cover, or some snow piles may look similar to snow layers over parked cars on the roadside. Programs trained for detecting cars and people in sunshine and snowless environments will fail to make sense of vehicles that are topped with piles of snow and people who are wearing several layers of clothing.

II. Related Works

The research on the perception of road environments has been rapidly progressing. Previous work such as [19] introduced a method to identify road surface types, such as snowy, icy, and wet, here by using a graininess analysis of acquired images with stereo image pairs. Another road status classification research [20] focused on sensors installed under the road surface, while some works [21], [22], and [23] have used a support vector machine (SVM) and the K-nearest neighbor (KNN) to determine road conditions. For road region detection, [24] and [25] recognized drivable areas of the road using a classical image processing technique based on vanishing point detection.

The semantic segmentation is a neural network that identifies every pixel in the image and classifies it into different classes. Moreover, various kinds of road surfaces support the neural network's working potential. Deep neural networks can solve some deficiencies of this method because it is a preferred method over other alternatives. Consequently, semantic segmentation was used based on various methods. A network called Fully Convolutional Networks (FCN) [26] was able to segment the image from any size by use of convolution neural networks (CNN) without fully connected layers making it a standard approach for new techniques. For instance, research [27] applied FCN to detect objects on a snowy road environment with the system accuracy, showing a mean intersection over union (mIoU) of about 50% with 7.84 FPS (frame per second). The FCN based network name DSNet [28], which maintained an accuracy from most previous ones, got 69.1% mIoU on the Cityscape dataset and 72.6% on the CamVid dataset. A real-time semantic segmentation [29] named ICNet used cascade feature fusion units to obtain segmentation, resulting in the Cityscape dataset was 69.5% mIoU with 30.3 FPS. An encoder-decoder-based developed network called Data-dependent Upsampling (DUpsampling) [30] was proposed to replace bilinear, which can recover the pixel-wise prediction from low-resolution outputs of CNNs. This model's main points were the improvement of reconstruction capability and flexibility of decoder in leveraging almost arbitrary combinations. Work on a slippery road caused by water, ice, and snow named D-UNet [31], developed from U-Net [32], used dilated convolutions for the sensible field of network. This technique got the highest performance as compared to classical machine learning. One

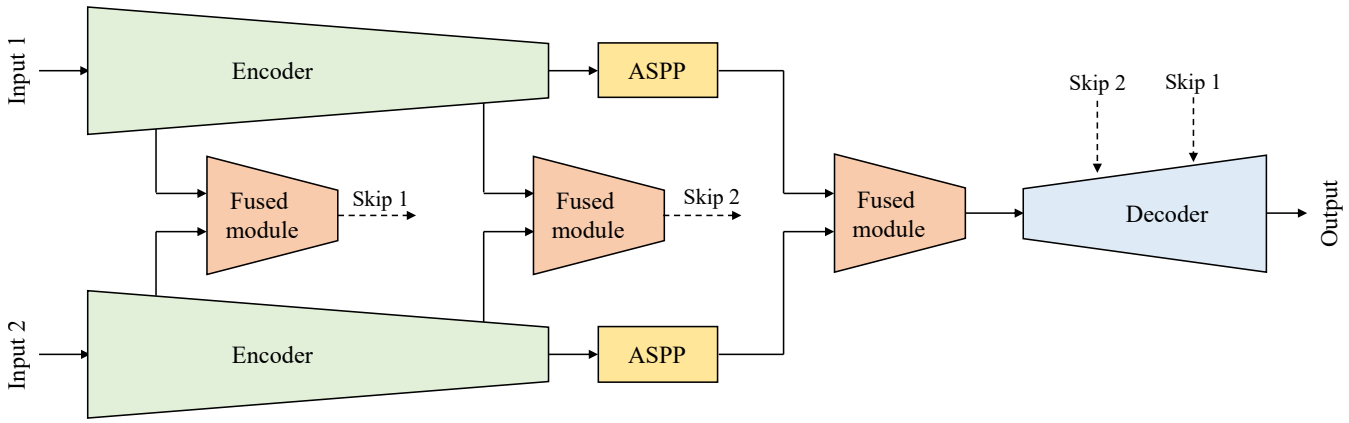


Fig. 2. Overall diagram of network architecture.

of the high-performance networks called Pyramid Scene Parsing Network (PSPNet) [33] was used for scene parsing of semantic segmentation. They applied the Pyramid Pooling Module, which consisted of four different pyramid scales to fuse the feature map, and the efficiency on the Cityscape dataset was about 82.6% mIoU. Gao et al. [34] developed an end-to-end framework from PSPNet called Multiple Feature Pyramid Network (MFPN) to function with road detection from satellite view, which further introduces a Tailored Pyramid Pooling Module to advance the accuracy of the original network. This model reached up to 7.8% mIoU higher than PSPNet with Massachusetts dataset. [35] proposed the Pointwise Spatial Attention Network (PSANet) to relieve the local neighborhood's constraint. The training with fine data and coarse+fine data were tested on the Cityscape dataset and got about 80.1% and 81.4% mIoU in sequence. A network name Dual Attention Network (DANet) [36] was developed to integrate local features with global dependencies. This model included two types of attention modules: the position attention module and the channel attention module. The position attention module combines each position feature by a weighted sum at all positions. The channel attention module emphasizes channel maps by adding features of all channel maps with a mIoU of 81.5% on the Cityscape dataset. Another significant research by Google Inc. [37] introduced a high-efficiency network called DeeplabV3, which applied Atrous Convolutions in parallel to extract the multi-scale context with different atrous rates. This model consisted of a 1×1 convolution filter, three 3×3 convolution filters with various atrous rates and a pooling feature. The efficiency on the Cityscape dataset was 81.3% mIoU. Some researchers [38] introduced their self-collected snowy scenario dataset to improve the neural network performance. The combination of the published dataset and their snowy dataset can improve the result of segmentation. Our previous work [39] tried to improve the training process for segmentation in a snowy environment. The network is based on DeeplabV3 with extracted auxiliary outputs to enhance the training performance.

A semantic segmentation technique can accommodate multiple inputs to enhance the segmentation's accuracy. For example, RedNet [40] improved the training procedure by fixing the gradient vanishing problems; this method was based on an encoder–decoder network operated with dual inputs, RGB, and depth images. It was able to calculate for four advanced outputs called the pyramid supervision training scheme in the decoder part. The encoder part of this network was based on the ResNet-50 model [41], with the decoder part operated by using the reversed ResNet-34 and ResNet-50 models, making this scheme achieve the best results on the SUN RGB-D dataset. Research on modal fusion for indoor environments [42] has proposed a model based on SIFT features and MRFs. Other works such as [43] have combined RGB and depth features using CNNs. This approach classifies pixels in the detection window as the foreground or background. The UpNet [44] network uses multispectral and multi-modal images for segmentation. The network contains fusion architecture that merges RGB, near-infrared channels, and depth information. Some researchers [45] introduced a network architecture consisting of two modality-specific encoder streams and a self-supervised model adaptation fusion module. Here, tests on the Cityscape, Synthia, and SUN RGB-D datasets achieved state-of-the-art performance compared with other networks. [46] also combined RGB with depth data for the SUN RGB-D dataset's indoor environment. The network architecture is an encoder–decoder type. The two encoder branches were used to extract the feature maps from both the RGB and depth data in parallel. Also, [47] introduced their network called depth-aware CNN. The two operations–depth-aware convolution and depth-aware average pooling were integrated into CNNs for segmentation in the RGB-D dataset. Not only was depth information used to improve the RGB segmentation, but the thermal information was also considered too. A work on an RGB-T dataset [48] introduced the use of RGB image and IR image segmentation for autonomous vehicles; the network was an encoder–decoder that extracted the features from two encoders in parallel and fused them in the decoder part. This network reached a higher accuracy than the other state-of-the-art segmentation methods. The RGB-thermal fusion network, or RTFNet, [49] was developed for segmentation in the urban scenario. Different from FuseNet [46] and MFNet [48], this

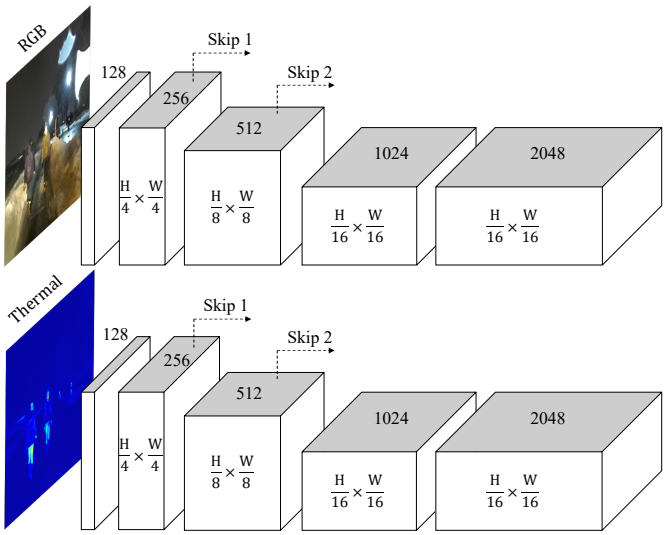


Fig. 3. Feature map shape in encoder processing.

network utilized ResNet [41] as the feature extractor. The RGB feature and thermal feature were fused in the RGB encoder part by element-wise addition. The Upception block has also been proposed as an initial part of the decoder. The works mentioned above describe multi-input networks that are accurate in good environment conditions, but they come up short in snowy scenarios.

Hence, in the current research, we study the semantic segmentation of roads in poor visibility environments particularly snowy conditions and snow-covered surfaces. We utilize multi-modal RGB-T semantic segmentation using the RGB images from an RGB camera and a thermal map from a thermal camera as the inputs of the network. This combination of information in snowy conditions provides excellent information about the subject. The network proposes a novel fusion module to combine two feature maps and use pyramid supervision to generate the side outputs to enhance training procedures.

III. Network Architecture

In this section, we describe the overall topology of our network. The proposed network uses a fully convolutional encoder–decoder pattern. Fig.2 shows the overall network architecture. The encoder follows the full pre-activation ResNet-50 model [41] because of the good balance between complex computations and deep feature learning. Our network contains two branches of the encoder for extracting features from the two inputs: RGB image and the thermal image. The Atrous spatial pyramid pooling (ASPP) module [37], the most efficient and state-of-the-art, is used to incorporate image-level features. Before passing the image to the decoder section, our ASPP module's outputs are merged by our fused module. The decoder part is also based on the ResNet-34 model but operates in the inverse direction to expand the image and reduce channels. This is adapted with the pyramid supervision training scheme of RedNet [40] to improve the training.

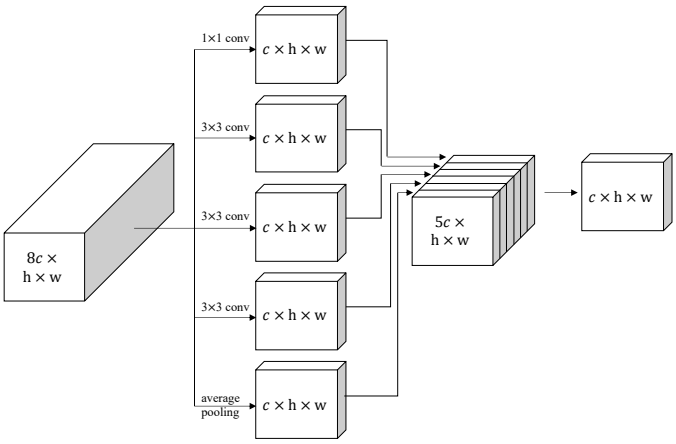


Fig. 4. Feature map shape in ASPP module.

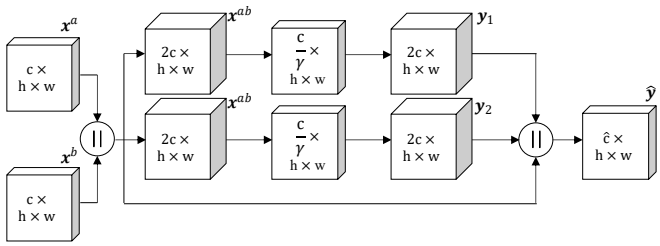


Fig. 5. Feature map shape in the fused module.

A. Encoder

This network's encoder has two similar convolution branches: the RGB branch and thermal branch. Both encoder branches are developed from the ResNet-50 model [41]; the adaptive average pooling layer and the linear transformation layer, which are the last two layers, are substituted with the ASPP module and our decoder. The same configuration is used in the RGB branch and thermal branch of this model. The encoder begins with a downsampling operation, a sequence of 7×7 convolution layers with stride two padding three, normalization, a rectified linear unit (ReLU) activation function, and 3×3 max-pooling layer with stride two and padding one. The next four layers of this encoder are residual layers with various numbers of residual units, and the residual number is $\{3, 4, 6, 3\}$. Fig. 3 shows the output shape of each layer, where H and W are the height and width of the input images, respectively. The original image is reduced to 1/16 times the original size, but the channel feature is increased to 2048 channels. During the encoder process, two skip feature maps are extracted between residual layers for the direct merger in the decoder process. The fusion module fuses the feature map of 256 channels after passing the first residual layer and the passed map of the second residual layer with 512 channels before feeding this into the decoder.

B. Atrous Spatial Pyramid Pooling

The ASPP module of DeeplabV3 [37] was developed from a previous version [50] by including batch normalization [51] into the model and adopting an image-level feature to incorporate the global context information. As the last

feature of this model, the global average pooling was applied and passed the output to a 1×1 convolution layer with 256 filters, the normalizer, and the linearize unit. As mentioned in Section 2, this module consists of one 1×1 convolution layer, 3×3 convolution layers with different rates, and a global average pooling layer in the image-level feature. In the current research, we use rates of $\{12, 24, 36\}$ for the 3×3 convolution layers in the module. The outputs from all branches are the same feature size at 256 channels, are concatenated, and are passed through to another 1×1 convolution layer with a normalizer to resize the channels back to the same size from each branch. Fig. 4 shows the output shape from each part of the module, where c, h, and w denote channel, height, and width of the feature map, respectively.

C. Fused module

For segmentation, only RGB images are not enough in the snowy environment because they contain only a feature of color. The differences in color cannot represent every object in the surrounding area, especially in snowy situations where snow reduces the gradient of different objects' colors, turning most of the area into the same color. Consequently, we introduced a thermal map that contains a temperature feature to support the loss feature from the snow. Thermal map utilization is more effective than the color image for separating living things from the ambient environment, which does not emit heat. Our network consists of two encoders, one for the RGB image and another for the thermal map, to improve snow segmentation. To merge together both feature maps from two encoders, we propose an architecture unit that can fuse and select the different maps' essential features. The network can learn to intensify the best informative features and minimize the less important information. Our fusion module is adapted from the AdapNet++ [45] fusion module, making it more suitable to work with a temperature feature map. The module begins with concatenating the two feature maps together and then feeds into three branches: a convolution branch (cb), a convolution with rate branch (cbr), and a skip branch (skip). Fig. 5 shows the diagram of our fusion module. Two convolution branches are a sequence of 3×3 convolution layers with a small number of filters, a linearized unit, a 3×3 convolution layer with the same number of filters that is a result of the concatenator, and the softplus function. This module introduces the softplus function as an element-wise function instead of ReLU in each branch's final step. Here, softplus is a smooth approximation to the ReLU function, as defined by (1), where β denotes the softplus constant.

$$\text{Softplus}(x) = \frac{1}{\beta}\log(1 + e^{\beta x}). \tag{1}$$

Given $\boldsymbol{x}^a \in \mathbb{R}^{c\times h\times w}$and $\boldsymbol{x}^b \in \mathbb{R}^{c\times h\times w}$ denote the feature maps from RGB images and the thermal map, where c is the channels of the feature and $h \times w$ is the dimension of the feature. The concatenated result of $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$ is represented with $\boldsymbol{x}^{ab} \in \mathbb{R}^{2c\times h\times w}$, and the results from each convolution branches are $\boldsymbol{y}_i \in \mathbb{R}^{2c\times h\times w}$, where $i \in \{1, 2\}$ is a branch number. The 3×3 convolution layers are represented by ϕ_ν^α, where α denotes the number of filters and ν is padding and dilation number. Equation (2) describes the operation in the convolution branch, where ρ is the rectified linear unit. In this case, we use a compression ratio γ of 32, padding ν of 1 for the first branch (cb) and 24 for another (cbr).

$$\boldsymbol{y}_i = \text{Softplus}\big[\phi_\nu^{\alpha_2}\rho\big(\phi_\nu^{\alpha_1}(\boldsymbol{x}^{ab})\big)\big]. \tag{2}$$

where the number of filters α can be calculated by (3)–(4), and γ denotes a compression ratio,

$$\alpha_1 = \frac{1}{\gamma}c, \tag{3}$$

$$\alpha_2 = 2c. \tag{4}$$

The results $\boldsymbol{y}_i$ are concatenated with $\boldsymbol{x}^{ab}$ before passing through the last convolution layer with $\widehat{c}$ filters padding of 1 and the normalizer ξ. Equation (5) describes the calculation for the final output $\widehat{\boldsymbol{y}}$ of this module, where $*$ is a concatenate operation.

$$\widehat{\boldsymbol{y}} = \xi\big[\phi_1^{\widehat{c}}(\boldsymbol{y}_1 * \boldsymbol{y}_2 * \boldsymbol{x}^{ab})\big]. \tag{5}$$

In this experiment, $\widehat{c}$ value of the fused module used with the outputs of the ASPP module is set to 1024. Therefore, the output of this fusion module has a feature size of 1024 channels with the same feature dimension as the module's input. Using this proposed module, we can obtain a merged output from two results of two encoder branches. For the fused module of the skip feature in the bypass section, the number of convolution filters in the last layers is set as equal to the module's input channels. Other networks from our model are trained with a concatenator as a fused module at the same position, and the skip feature maps are not operated in these networks.

D. Decoder

The decoder of this model is developed from the ResNet-34 model [41] but operate in the inverse direction of the original procedure. The decoder begins with the upsample operation, which is the last block of ResNet-34. Unlike the encoder, a sequence of a 7×7 convolution layer and 3×3 max-pooling layer are removed at the last convolution layer of each layer. A transposed convolution operator substitutes for the convolution filter to double the size of the dimension. The size of the feature map from each layer can be calculated by (6),

$$size_{out} = stride(size_{in} - 1) + n - 2r + out_{padding}. \tag{6}$$

where $size_{in,out}$ are the size of the feature map, height, or width, respectively; stride is the $stride$ of the convolution, n denotes the size of the convolving kernel, r denotes the size of zero-padding that is added to both sides of each dimension in the input, and $out_{padding}$ is an additional size added to one side of each dimension in the output size. Fig. 6 shows a comparison of the bottleneck between the encoder and decoder.

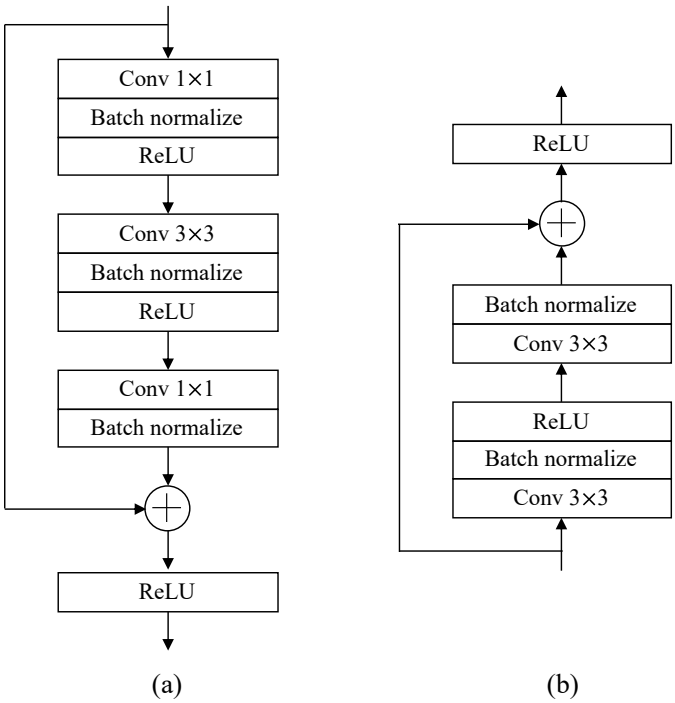


Fig. 6. Bottleneck of (a) encoder and (b) decoder.

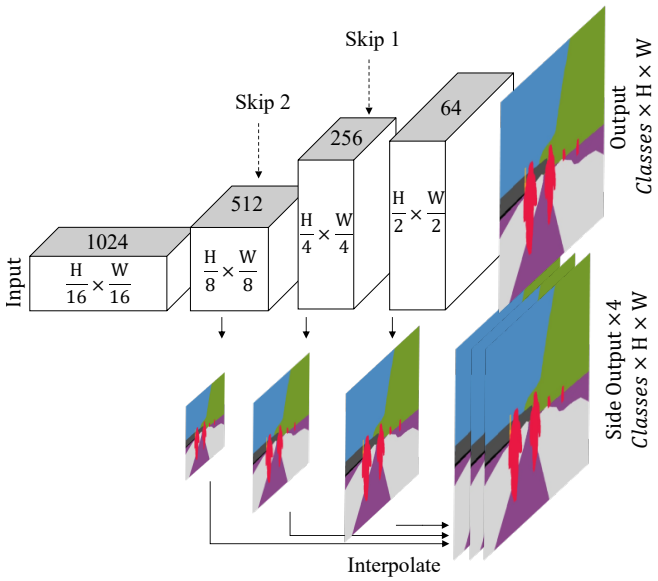


Fig. 7. Feature map shape in decoder processing.

The number of layers is also similar to the encoder because of the input channels of the encoder, and the number of bottlenecks for each layer is $\{6, 4, 3, 3\}$. This decoder ends with a transposed convolution operator, which reduces the channel number into design classes and doubles the size of the dimension to the same as the input of the network. Based on the pyramid supervision of RedNet [40], the final output is generated, and the other three outputs from each upsample layer are extracted. These extra outputs are called side outputs and will be used to calculate losses together with the final output. The output of each layer is passed through a 1×1 convolution with the same number of filters to fix classes and

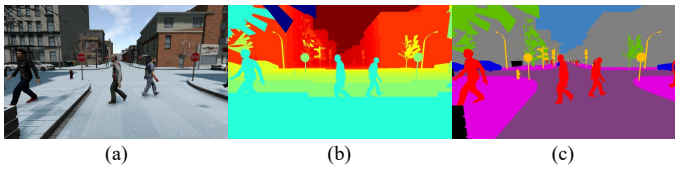


Fig. 8. Example images of Synthia dataset, (a) RGB image, (b) depth image, and (c) groundtruth image.

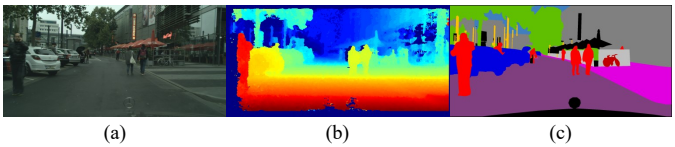


Fig. 9. Example images of Cityscape dataset, (a) RGB image (b) depth image, and (c) groundtruth image.

then will be interpolated with bilinear interpolation function, as shown by (7)–(8), to the original size.

$$f(x, y) = a_0 + a_1 + a_2 y + a_3 xy, \tag{7}$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_1 & y_2 & x_1 y_2 \\ 1 & x_2 & y_1 & x_2 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \end{bmatrix}^{-1} \begin{bmatrix} f(x_1, y_1) \\ f(x_1, y_2) \\ f(x_2, y_1) \\ f(x_2, y_2) \end{bmatrix}. \tag{8}$$

where $f(x, y)$ is a value of the unknown function at the point (x, y) and a_0 to a_3 are the outputs that are passed through the cross-entropy function. The values of $x_{1,2}$ and $y_{1,2}$ are the known values of the four points. The side outputs are created only in the training mode to improve the training process but are inactivated in evaluation mode. Fig. 7 shows the output diagram of the decoder.

IV. Experiment

This section introduces the dataset used in our experiment and the training environments with experimental results. The evaluation method and comparative analysis are also explained in detail.

A. Dataset

Our network is evaluated on two public datasets that are available for testing urban driving scenarios and on another dataset collected for the current study for the application of a snow removal machine in snowy road environments. The selected datasets combine challenging conditions for perception, including snow, slush, motion-blur, night, and so forth. Each dataset contains multiple input images for training suitable to test the fusion module of our network. The three datasets used in this research are the Synthia dataset [52], Cityscapes dataset [53], and snowy road dataset (new).

The Synthia dataset [52] is an outdoor dataset using the Unity engine [54] to render a virtual city. It contains 4686 sets of realistic images, depth information, and annotated label sets that have an image resolution of 1280×760. The current research uses only the winter datasets of all packages,

TABLE I
DATASETS INFORMATION.

ID	Name	pixels (pixel)	pixels (%)	Color
	Synthia			
1	Road	6.3e+8	13.7	
2	Sidewalk	9.2e+8	20.1	
3	Building	1.8e+9	40.4	
4	Fence	1.8e+8	3.9	
5	Traffic object	1.1e+8	2.4	
6	Vegetation	1.4e+8	3.0	
7	Sky	5.6e+7	1.2	
8	Pedestrian	1.8e+8	4.0	
9	Vehicle	1.1e+7	0.2	
10	Unidentified	5.0e+8	10.9	
	Cityscape			
1	Road	2.4e+9	33.4	
2	Sidewalk	3.9e+8	5.3	
3	Building	1.5e+9	20.3	
4	Fence	9.9e+7	1.4	
5	Traffic object	1.3e+8	1.8	
6	Vegetation	1.1e+9	15.2	
7	Sky	2.5e+8	3.5	
8	Pedestrian	8.9e+7	1.2	
9	Vehicle	5.0e+8	6.9	
10	Bicycle	2.9e+7	0.4	
11	Unidentified	7.7e+8	10.5	
	Snowy Road			
1	Road	6.0e+7	17.5	
2	Snow mountain	8.7e+7	25.3	
3	Building	1.8e+7	5.3	
4	Traffic object	1.0e+6	0.3	
5	Vegetation	8.0e+7	23.0	
6	Sky	8.2e+7	23.9	
7	Pedestrian	7.1e+6	2.1	
8	Vehicle	2.0e+6	0.6	
9	Unidentified	7.1e+6	2.1	

including Highway Seqs-01, New York ish Seqs-02, Old European Town Seqs-04, New York ish Seqs-05, and Highway Seqs-06. The classes of this dataset are reduced to 10 classes to suit our specific goal needs, as shown in Table I. Examples of the Synthia dataset are shown in Fig. 8.

The Cityscapes dataset [53] contains real RGB-D images for road and city environments; it consists of complex surroundings with various weather conditions of more than 50 cities. The dataset was collected by a stereo camera with a resolution of 2048×1024 pixels, so it contains RGB images and disparity images. The left 8-bit images set of this Cityscape dataset are used to evaluate the network. This dataset contains 3474 images of a training, validation, and testing set that are labeled into 11 classes, as shown in Table I. Fig. 9 shows example images of the Cityscapes dataset.

The snowy road dataset is a new dataset collected for the current study and for the snowy road environment classification in the Hokkaido region to improve the operation of snow removal machines. The environment contains 2458 images collected in both the day and night and in snowy conditions. It contains images of urban roads and sidewalks covered with snow, along with pedestrians. The images are labeled into nine classes: road, snow mountain, building, traffic object, vegetation, sky, human, vehicle, and background. Table I shows the details of the dataset. This dataset was collected by the RGB camera and Optris thermal camera with a resolution of 450×350 pixels. Fig. 10 shows some example images from our dataset. This dataset is separated into two sub-datasets because of the different conditions of the environment: the SRM and SSW dataset. The snow removal machine (SRM) sub-dataset contains both the day and night environment of snowy public roads that require snow removal operations, as shown in the bottom row of Fig.10. There are many snow-covered mountains on both sides of the road and many pedestrians walking across the street during machine operation. This sub-dataset consists of 1571 images. The snowy sidewalk (SSW) sub-dataset contains only day conditions of snowy sidewalks and surrounding in Hokkaido, as shown in the top row of Fig.10. This environment is the sidewalk on the side of the street. There are snowy mountains only on one side of the sidewalk. The pedestrians in this scenario, who are walking along the sidewalk, walk close to and farther from the machine.

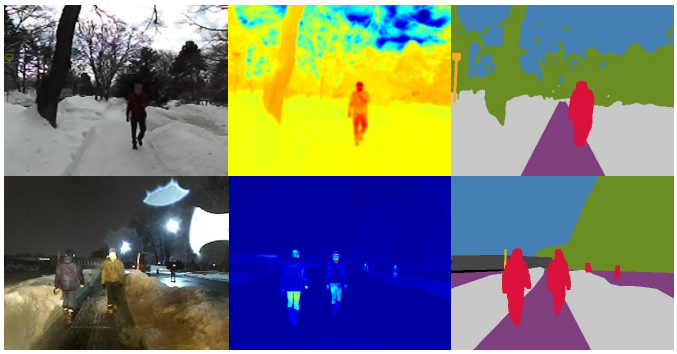


Fig. 10. Example images of SRM dataset, (a) RGB images, (b) thermal images, and (c) groundtruth images.

B. Training Environment

In the training process, all RGB images, thermal maps (images), and groundtruth maps are resized into 480 height and 640 width. A bilinear interpolation is used for the input images, but the nearest-neighbor interpolation is applied for the groundtruth. Additionally, thermal maps, which are one-channel images, are converted into three-channel images of the jet colormap. The experiment is processed on a single GPU of NVIDIA Geforce GTX 1080Ti and CPU of Intel i7-4790 with 16 GB of memory. The deep learning framework Pytorch is used to build and train the network model. The network encoder is initialized by pre-training on the ImageNet dataset [55], but other layers' parameters are randomly initialized. The training loss of each class is reweighted by cross-entropy function as (9), where m is the number of classes, p is the predicted probability observation, and q is the binary indicator (0 or 1).

TABLE II
THE COMPLEXITY COMPARISON OF NETWORKS.

Network	Dataset RGB	Dataset T	Parameters	Memory	MACs	Time
ICNet [29]	✓	-	28.29M	0.93G	18.7G	52ms
	✓	✓	30.73M	1.15G	36.65G	112ms
DANet [36]	✓	-	49.62M	1.78G	75.3G	31ms
	✓	✓	49.63M	2.46G	150.6G	63ms
DUpsampling [30]	✓	-	34.53M	1.49G	57.19G	28ms
	✓	✓	39.47M	2.35G	109.78G	58ms
PSANet [35]	✓	-	52.97M	2.19G	144.85G	31ms
	✓	✓	71.85M	3.81G	289.67G	61ms
PSPNet [33]	✓	-	48.76M	1.69G	69.29G	31ms
	✓	✓	67.63M	2.79G	138.58G	59ms
DeeplabV3 [37]	✓	-	41.81M	1.46G	65.3G	31ms
	✓	✓	42.4M	2.21G	130.6G	59ms
DeeplabV3 [37] + SSMA [45]	✓	✓	42.54M	2.53G	130.77G	60ms
Ours	✓	✓	45.5M	2.27G	134.31G	58ms
Ours + pyout	✓	✓	109M	3.53G	393.29G	75ms

M is $\times 10^6$, G is $\times 10^9$

$$Loss = -\frac{1}{m}\sum_{m} q\log(p) + (1-q)\log(1-p), \tag{9}$$

$$Losses = \sum_{i} loss_i. \tag{10}$$

The losses of all outputs are summed before differentiating the losses as (10), where i is the number of outputs. The optimization algorithm is momentum SGD, which is set to 0.9, and a weight decay of 0.0001 is used. The initial learning rate ($rate_{ini}$) is set to 0.001 and decay 0.95 at every 50 epochs, and the learning rate (lr) is calculated by (11), where $epochs$ denotes the global step epochs of training and dpe is the decay per epochs.

$$lr = rate_{ini} \times decay^{\frac{epochs}{dpe}}. \tag{11}$$

The complexity of our network is measured and compared with other competitive networks. The increasing number of parameters, GPU memory, MACs(multiply-accumulate), and processing time are used as the indicators of this experiment. This test is done with a single batch size of 3×640×480 input image. Table II shows the complexity comparison of the networks. Our network with pyramid output has the highest number of parameters but utilizes lesser GPU memory than PSANet [35]. The processing speed of our network is also faster than ICNet [29] when feeding a single batch size of the input.

V. RESULTS & EVALUATION

The proposed network is evaluated on the SRM and SSW datasets as the primary dataset for the snowy environment and the other two datasets, the Cityscape and the Synthia, to support the results. Because the current research's primary purpose is to perform segmentation in a snowy road environment, drivers and pedestrian safety become the top priority. Therefore, there are three criteria used in this experiment to determine the efficacy of our network, that is, the mean accuracy (mAcc), mIoU, and mean dice coefficient (mF1). The accuracy is the percent of pixels in the results that are classified correctly, and the calculation is shown by (12). The intersection over union is an essential evaluation method for segmentation, and (13) describes the computation of IoU. The F1 score is the harmonic mean of the precision and recall. It is calculated by (14), where TP is true-positive, TN is true-negative, FP is false-positive, and FN is false-negative.

TABLE III
THE MIOU OF EACH COMBINATION ON THE SRM DATASET.

Padding number (ν)	Compression ratio (γ)	mIoU(%)
6	32	70.9
12	32	72.6
18	32	71.2
24	**32**	**73.2**
30	32	71.8
36	32	72.0
24	2	72.5
24	4	72.3
24	8	71.1
24	16	70.4
24	**32**	**73.2**

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, \tag{12}$$

$$IoU = \frac{TP}{TP+FP+FN}, \tag{13}$$

$$F1 = \frac{2TP}{2TP+FP+FN}. \tag{14}$$

TABLE IV
THE ABLATION TEST OF THE FUSED MODULE ON THE SRM DATASET.

Fused module architecture	Human IoU (%)	mIoU (%)
cb + cbr + skip	**75.8**	**73.2**
cb + cbr	73.8	71.9
cb + skip	73.5	72.3
cbr + skip	74.0	72.0
cb	74.0	71.3
cbr	72.7	71.4
skip	73.6	72.3

A. Combination of the compression ratio and the padding number

There are two main parameters in the fused module that can affect the efficiency of the network: the compression ratio (γ) is used to reduce the size of the channels in each branch, and the padding number (ν) is used to capture the multi-scale context in a feature map. Our fused module is tested for the most suitable combination of γ and ν in a snowy environment. The values of the compression ratio are varied by a power of two and multiple of six for the padding and dilation value. Table III shows the mIoU of each combination of γ and ν on the SRM dataset. The results show that the best combination for a snowy environment is $\gamma = 32$ and $\nu = 24$.

B. The ablation study of the fused module

The ablation study of the proposed fused module is done on the SRM dataset. The network analyzed in this experiment consists of the encoder, ASPP module, and fused module. This study is tested on seven combinations of the fused module: cb, cbr, skip, cb+cbr, cb+skip, cbr+skip, and cb+cbr+skip (our proposal). The IoU of the pedestrian class and mean IoU are used as the indicators to evaluate the module. Table IV shows the results of the ablation test of the fused module on the SRM dataset. The results show that using double branches (cb+cbr) is better than only a single branch (cb or cbr) for the mean IoU, and the utilization of a skip branch (skip) can improve the pedestrian IoU and mean IoU of each convolution branch. The whole combination of each fused module branch reaches the highest of every condition in both the pedestrian's IoU and mean IoU.

C. Comparison of RGB and RGB-T on pedestrian IoU

The thermal map provides excellent information on heat reflected from objects, especially when used for discovering humans. In a snowy environment, the temperature difference between the surrounding environment and a human can be clearly represented in the thermal map. The further use of this heat information can improve human detection efficiency. The efficiency of using the thermal map as a second input is described in Table V. This experiment is the comparison between using the RGB input *with* and *without* thermal map input.

The test is focused on the IoU of humans, which is called human detection, and tested on many state-of-the-art networks.

TABLE V
THE COMPARISON OF USING THE THERMAL MAP IN THE SRM DATASET.

Network	Human IoU (%)	
	RGB	RGB-T
ICNet [29]	57.6	66.1
DANet [36]	67.2	72.4
DUpsampling [30]	2.3	64.2
PSANet [35]	68.4	58.9
PSPNet [33]	64.5	74.1
DeeplabV3 [37]	64.9	72.9
DeeplabV3 [37] + SSMA [45]	-	74.8
Ours	-	**75.8**
Ours + pyout	-	**78.4**

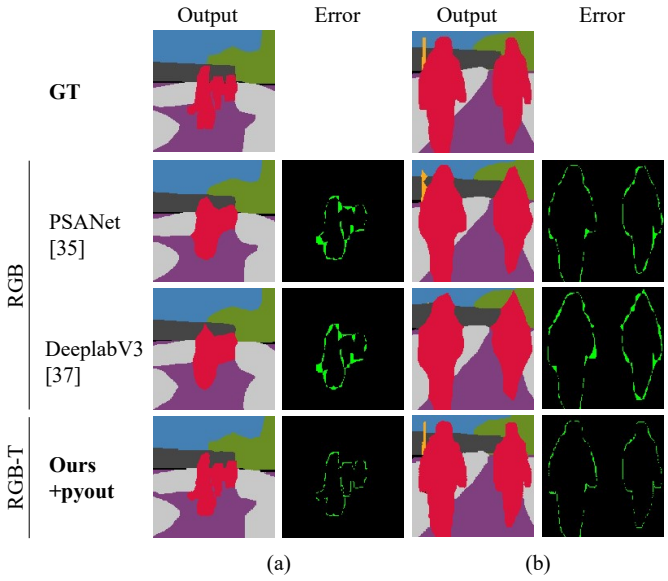


Fig. 11. Segmentation errors in human detection, (a) Daytime environment, (b) Nighttime environment.

The results show that using a thermal map can improve the efficiency of human detection for most networks. The proposed fused module can reach the highest IoU in the human class, and the application of the decoder can be higher than the previous one. Fig. 11 shows examples of segmentation errors in human detection. The green area is the pixel difference of the segmentation results from ground truth (GT). Our network using the RGB-T inputs obtained a smaller error for human detection than the best state-of-the-art network using only RGB images as the input.

D. Comparison of the proposed network and other state-of-the-art networks

Our network was tested and compared with other state-of-the-art networks. Table VI shows the IoU of each class on the RGB-T of the SRM and SSW datasets. The results show that our network, which uses fused modules, can obtain a higher mIoU and mF1 than other networks, and with the inverse ResNet-34 decoder, it can reach the highest IoU in all classes. The results show that our network becomes the

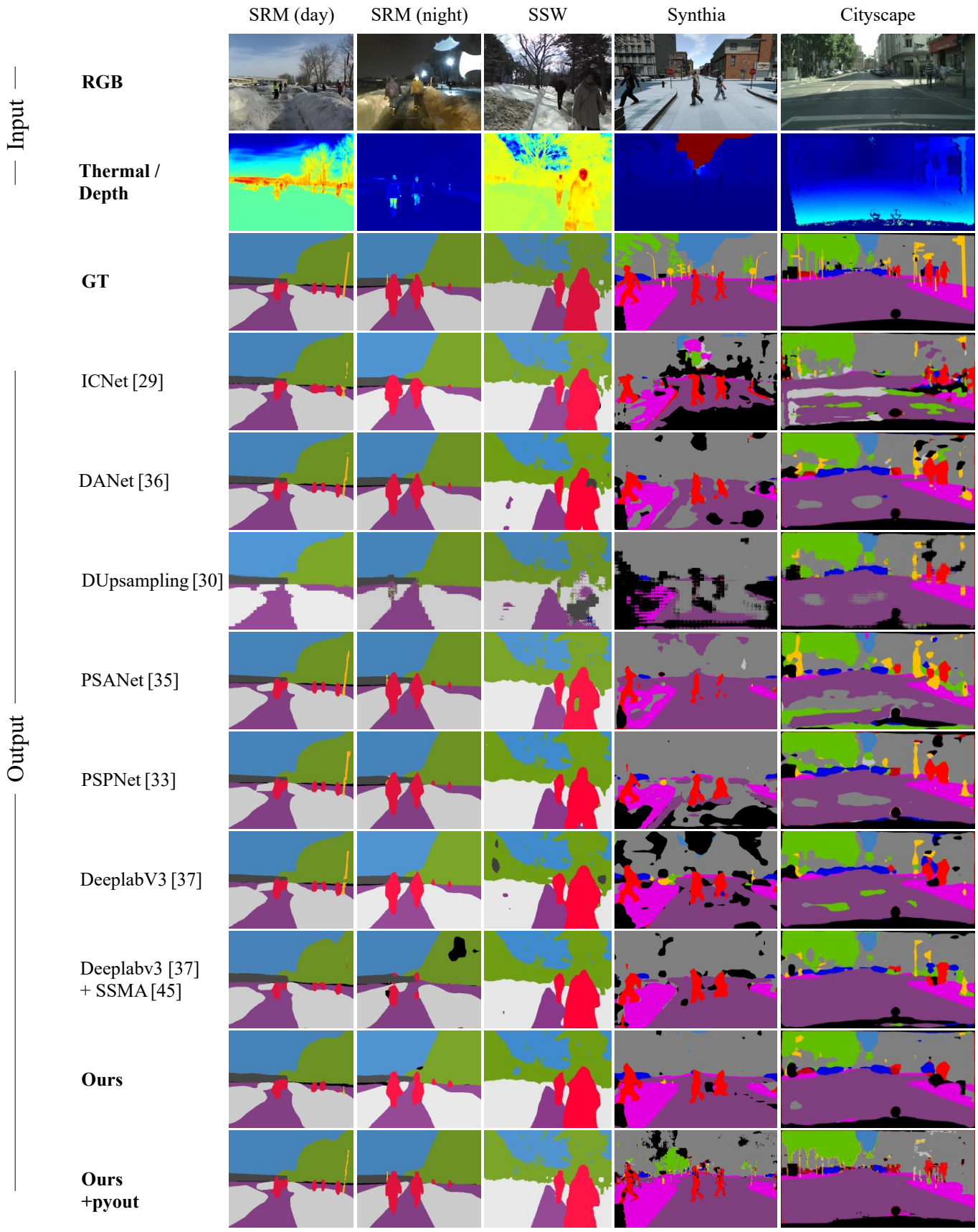


Fig. 12. Results of segmentation on all datasets.

TABLE VI
THE SEGMENTATION PERFORMANCE ON THE SRM DATASET AND SSW DATASET.

ID	IoU (%) 1	2	3	4	5	6	7	8	9	mIoU(%)	mF1(%)
SRM dataset											
ICNet [29]	83.2	77.0	74.9	24.1	86.6	54.1	66.1	11.7	22.4	55.6	67.0
DANet [36]	85.4	91.3	85.7	37.7	87.7	91.2	72.4	18.9	35.9	67.4	76.9
DUpsampling [30]	84.2	90.4	81.0	0.1	87.0	93.7	64.2	0.9	39.9	60.2	67.0
PSANet [35]	82.4	88.5	80.3	38.7	87.9	91.8	58.9	16.2	47.2	65.8	76.1
PSPNet [33]	87.3	89.8	83.9	34.6	90.1	88.9	74.1	20.7	52.6	69.1	78.6
DeeplabV3 [37]	85.0	90.6	82.6	33.4	93.2	95.0	72.9	19.8	59.4	70.2	79.3
DeeplabV3 [37] + SSMA [45]	85.7	91.4	83.9	36.7	92.5	95.3	74.8	26.6	53.7	71.2	80.4
Ours	**87.5**	**92.7**	**85.4**	**36.4**	**94.4**	**95.9**	**75.8**	**27.5**	**63.3**	**73.2**	**81.9**
Ours + pyout	**91.8**	**95.5**	**91.4**	**53.5**	**95.6**	**96.5**	**81.8**	**31.9**	**67.4**	**78.4**	**86.5**
SSW dataset											
ICNet [29]	89.4	87.4	88.3	8.2	89.5	92.7	79.7	79.7	24.9	71.1	78.3
DANet [36]	88.2	86.9	87.2	16.1	86.4	88.1	82.8	78.9	26.2	71.2	79.4
DUpsampling [30]	87.3	83.7	80.2	0.5	86.9	90.5	1.5	22.2	0.0	50.3	55.7
PSANet [35]	89.0	87.4	87.5	18.0	87.7	89.7	85.8	70.2	19.4	70.5	78.5
PSPNet [33]	89.5	87.5	88.5	18.8	86.7	87.8	86.1	83.3	32.1	73.4	81.3
DeeplabV3 [37]	88.5	86.6	81.6	17.2	83.6	88.7	85.5	81.8	30.6	71.6	80.0
DeeplabV3 [37] + SSMA [45]	91.9	90.5	89.5	14.1	87.7	89.5	84.5	74.7	21.3	71.5	79.8
Ours	**88.0**	**86.5**	**88.5**	**18.2**	**87.0**	**89.2**	**88.9**	**85.3**	**32.2**	**73.8**	**81.5**
Ours + pyout	**94.1**	**93.5**	**92.2**	**33.8**	**93.8**	**95.3**	**93.7**	**84.5**	**49.1**	**81.1**	**87.6**

state-of-the-art for segmentation in snowy environments. Fig. 12 shows the results of segmentation in every dataset.

To ensure the efficiency of the fused module, we evaluated the proposed multi-modal fused network on other published datasets and tested for mAcc, mIoU, and mF1. Table VII shows the results of mAcc, mIoU, and mF1 on the other two datasets, the Synthia and the Cityscapes. The results show that our fused module can improve the efficiency of multiple inputs of the neural network and train using a pyramid supervision path to generate the highest values for every indicator.

VI. CONCLUSION

In a snowy road environment during the day, visibility is hampered because of reflection from snow on the road. Also, for snow removal machines operating in the night, it is very difficult to recognize objects in the vicinity. Traditional neural networks that produce a high accuracy in clear snow-less environment do not function well in snowy environments. Our network has been developed to work in this inclement weather condition. During nighttime snow removal machine operations, the most important objective is to detect humans. The proposed network can respond to this class very well because the input of the network includes thermal information, so it is more accurate than other popular networks.

In conclusion, the current work has introduced a fused module for multi-input neural networks to use in snowy road environments. In a snowy environment, the use of a thermal map as a second input is better than a single input of an RGB image. The proposed method shows robust human detection and higher mIoU in the snowy dataset. The network with the fused module and pyramid supervision path reached up to 78% mIoU and has become the state-of-the-art network for snowy environments. From the results on the Synthia and Cityscape datasets, we confidently can say that our fused module can enhance the efficiency of multiple input networks. The combination of our fused module and pyramid supervision path obtained the best results in both mAcc and mIoU. Currently, the proposed network is being developed for use with snow removal machine operation where there are sidewalks and snow road conditions. The network is trained with only nine classes of the objects. The datasets do not include every object in daily life, for example, other heat-emitting objects like animals or maintenance holes, which can also be represented in the thermal image. For future work, we plan to apply other information from LiDAR, RADAR, or stereo cameras to create an RGB-D-T dataset to improve our model's efficiency in snowy environments.

REFERENCES

[1] D. Eisenberg and K. E. Warner, "Effects of snowfalls on motor vehicle collisions, injuries, and fatalities," *American Journal of Public Health*, vol. 95, no. 1, pp. 120–124, 2005.

[2] J. Weng, L. Liu, and J. Rong, "Impacts of Snowy Weather Conditions on Expressway Traffic Flow Characteristics," *Discrete Dynamics in Nature and Society*, vol. 2013, p. 791743, 2013.

[3] F. Malin, I. Norros, and S. Innamaa, "Accident risk of road and weather conditions on different road types," *Accident Analysis and Prevention*, vol. 122, no. February 2018, pp. 181–188, 2019.

[4] K. Sano, T. Inagaki, J. Nakano, and N. C. Y, "An Analysis on Traffic Accidents on Undivided Expressway in Cold and Snow Area," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 8, pp. 2048–2061, 2010.

[5] A. A. Ravankar, Y. Hoshino, A. Ravankar, L. Jixin, T. Emaru, and Y. Kobayashi, "Algorithms and a framework for indoor robot mapping in a noisy environment using clustering in spatial and hough domains," *International Journal of Advanced Robotic Systems*, vol. 12, no. 3, p. 27, 2015.

TABLE VII
THE SEGMENTATION PERFORMANCE ON THE CITYSCAPE DATASET AND SYNTHIA DATASET.

Network	mIoU(%)	mAcc(%)	mF1(%)
		Cityscape	
ICNet [29]	40.0	96.0	51.7
DANet [36]	58.6	97.7	70.9
DUpsampling [30]	53.1	97.4	65.3
PSANet [35]	48.3	96.0	62.7
PSPNet [33]	56.1	97.5	68.6
DeeplabV3 [37]	59.8	97.7	72.3
DeeplabV3 [37] + SSMA [45]	58.2	97.5	71.4
Ours	**59.6**	**97.7**	**73.2**
Ours + pyout	**62.6**	**98.1**	**73.8**
		Synthia	
ICNet [29]	27.9	89.5	36.5
DANet [36]	51.3	94.0	64.2
DUpsampling [30]	33.0	91.0	44.1
PSANet [35]	55.0	97.2	67.9
PSPNet [33]	47.1	92.0	59.7
DeeplabV3 [37]	62.4	97.1	73.9
DeeplabV3 [37] + SSMA [45]	61.5	98.0	72.5
Ours	**65.1**	**98.1**	**76.1**
Ours + pyout	**69.4**	**98.7**	**78.5**

[6] A. Ravankar, A. A. Ravankar, Y. Kobayashi, Y. Hoshino, and C.-C. Peng, "Path smoothing techniques in robot navigation: State-of-the-art, current and future challenges," *Sensors*, vol. 18, no. 9, 2018.

[7] Z. Chen, J. Zhang, and D. Tao, "Progressive lidar adaptation for road detection," *CoRR*, vol. abs/1904.01206, 2019.

[8] F. Xu, L. Chen, J. Lou, and M. Ren, "A real-time road detection method based on reorganized lidar data," *PLOS ONE*, vol. 14, pp. 1–17, 04 2019.

[9] M. Marchetti, M. Moutton, S. Ludwig, L. Ibos, V. Feuillet, and J. Dumoulin, "Implementation of an infrared camera for road thermal mapping," in *10th International Conference on Quantitative InfraRed Thermography*, 2010.

[10] Y. Li, W. Ding, X. Zhang, and Z. Ju, "Road detection algorithm for autonomous navigation systems based on dark channel prior and vanishing point in complex road scenes," *Robotics and Autonomous Systems*, vol. 85, pp. 1 – 11, 2016.

[11] C. Fernández, D. Fernández-Llorca, and M. A. Sotelo, "A Hybrid Vision-Map Method for Urban Road Detection," *Journal of Advanced Transportation*, vol. 2017, 2017.

[12] S. Vachmanus, T. Emaru, A. Ravankar, and Y. Kobayashi, "Road detection in snowy forest environment using rgb camera," in *Proceedings of the 36 Annual Conference of the RSJ*, 2018.

[13] J. Baili, M. Marzougui, A. Sboui, S. Lahouar, M. Hergli, J. S. C. Bose, and K. Besbes, "Lane departure detection using image processing techniques," in *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, pp. 238–241, 2017.

[14] G. Somasundaram, "Lane Change Detection and Tracking for a Safe-Lane Approach in Real Time Vision Based Navigation Systems," pp. 345–361, 2011.

[15] X. Jiang, B. Hu, S. Chandra Satapathy, S.-H. Wang, and Y.-D. Zhang, "Fingerspelling Identification for Chinese Sign Language via AlexNet-Based Transfer Learning and Adam Optimizer," *Scientific Programming*, vol. 2020, p. 3291426, 2020.

[16] X. Jiang, S. C. Satapathy, L. Yang, S.-H. Wang, and Y.-D. Zhang, "A Survey on Artificial Intelligence in Chinese Sign Language Recognition," *Arabian Journal for Science and Engineering*, vol. 45, no. 12, pp. 9859–9894, 2020.

[17] P. S. Zaki, M. M. William, B. K. Soliman, K. G. Alexsan, K. Khalil, and M. El-Moursy, "Traffic signs detection and recognition system using deep learning," 2020.

[18] M. Thoma, "A survey of semantic segmentation," *CoRR*, vol. abs/1602.06541, 2016.

[19] M. Jokela, M. Kutila, and L. Le, "Road condition monitoring system based on a stereo camera," in *2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing*, pp. 423–428, aug 2009.

[20] A. Troiano, E. Pasero, and L. Mesin, "New system for detecting road ice formation," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 3, pp. 1091–1101, 2011.

[21] R. Omer and L. Fu, "An automatic image recognition system for winter road surface condition classification," in *13th International IEEE Conference on Intelligent Transportation Systems*, pp. 1375–1379, 2010.

[22] P. Jonsson, J. Casselgren, and B. Thornberg, "Road surface status classification using spectral analysis of NIR camera images," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1641–1656, 2015.

[23] S. Kawai, K. Takeuchi, K. Shibata, and Y. Horita, "A smart method to distinguish road surface conditions at night-time using a car-mounted camera," *IEEJ Transactions on Electronics, Information and Systems*, vol. 134, no. 6, pp. 878–884, 2014.

[24] N. John, B. Anusha, and K. Kutty, "A Reliable Method for Detecting Road Regions from a Single Image Based on Color Distribution and Vanishing Point Location," *Procedia Computer Science*, vol. 58, pp. 2–9, 2015.

[25] H. Kong, J. Y. Audibert, and J. Ponce, "General road detection from a single image," *IEEE Transactions on Image Processing*, vol. 19, no. 8, pp. 2211–2220, 2010.

[26] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[27] Z. Pan, T. Emaru, A. Ravankar, and Y. Kobayashi, "Applying semantic segmentation to autonomous cars in the snowy environment," *arXiv preprint arXiv:2007.12869*, 2020.

[28] P.-R. Chen, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "DSNet: An Efficient {CNN} for Road Scene Segmentation," *CoRR*, vol. abs/1904.0, 2019.

[29] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," *CoRR*, vol. abs/1704.0, 2017.

[30] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 3121–3130, 2019.

[31] C. Liang, J. Ge, W. Zhang, K. Gui, F. A. Cheikh, and L. Ye, "Winter Road Surface Status Recognition Using Deep Semantic Segmentation Network," in *International Workshop on Atmospheric Icing of Structures*, (Reykjavík), pp. 1–6, 2019.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.0, 2015.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6230–6239, 2017.

[34] X. Gao, X. Sun, Y. Zhang, M. Yan, G. Xu, H. Sun, J. Jiao, and K. Fu, "An End-to-End Neural Network for Road Extraction from Remote Sensing Imagery by Multiple Feature Pyramid Network," *IEEE Access*, vol. 6, pp. 39401–39414, 2018.

[35] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise Spatial Attention Network for Scene Parsing," in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 270–286, Springer International Publishing, 2018.

[36] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual Attention Network for Scene Segmentation," *CoRR*, vol. abs/1809.0, 2018.

[37] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *CoRR*, vol. abs/1706.0, 2017.

[38] Y. Lei, T. Emaru, A. A. Ravankar, Y. Kobayashi, and S. Wang, "Semantic image segmentation on snow driving scenarios," in *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1094–1100, 2020.

[39] S. Vachmanus, A. A.Ravankar, T. Emaru, and Y. Kobayashi, "Semantic segmentation for road surface detection in snowy environment," in *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 1381–1386, 2020.

[40] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation," *CoRR*, vol. abs/1806.0, 2018.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.0, 2015.

[42] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 601–608, 2011.
[43] S. Gupta, R. B. Girshick, P. Arbelaez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," *CoRR*, vol. abs/1407.5, 2014.
[44] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion," in *International Symposium on Experimental Robotics (ISER 2016)*, pp. 465–477.
[45] A. Valada, R. Mohan, and W. Burgard, "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1239–1285, 2020.
[46] C. Hazırbaş, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," 11 2016.
[47] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," *CoRR*, vol. abs/1803.06791, 2018.
[48] H. Qishen, W. Kohei, K. Takumi, U. Yoshitaka, and H. Tatsuya, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115, 2017.
[49] S. Yuxiang, Z. Weixun, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.
[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *CoRR*, vol. abs/1606.0, 2016.
[51] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *CoRR*, vol. abs/1502.0, 2015.
[52] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, no. 600388, pp. 3234–3243, 2016.
[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *CoRR*, vol. abs/1604.0, 2016.
[54] J. K. Haas, "A history of the unity game engine," 2014.
[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

Sirawich Vachmanus received his B.E. degree in Mechatronics Engineering from King Mongkut's University of Technology Thonburi, Thailand, in 2016, and M.E. degree in Human Mechanical System and Design Engineering from Hokkaido University, Japan, in 2019. He is currently a doctoral course student at the research laboratory of Robotics and Dynamics at the graduate School of Engineering, Hokkaido University. His research interests include computer vision, deep learning, sensor fusion, and artificial intelligence for robots. He is a student member of the IEEE.

Ankit A. Ravankar received his M.E. and PhD degrees from Hokkaido University, Japan, in 2012 and 2015, respectively. He was the recipient of the MEXT scholarship from the government of Japan for his graduate studies. From 2015-2021 he was an Assistant Professor at the Faculty of Engineering, Division of Human Mechanical Systems and Design, Hokkaido University, Japan. Since April 2021 he is working as a Research Associate/ Lecturer at the Division of Mechanical Engineering, Department of Robotics, Tohoku University, Japan. His research interests include planning and decision making for robot sensing under uncertainty, mobile robot navigation, autonomous vehicles, simultaneous localization and mapping (SLAM), path planning, computer vision, and multi-robot systems. He is a member of the IEEE, SICE, and JSME.

Takanori Emaru received his M.E. and PhD degrees in Electrical Engineering from Hokkaido University, Japan, in 1998 and 2002, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science at the University of Electro-Communications, Japan, from 2003 to 2006. He was a Lecturer at Osaka Electro-Communication University from 2006 to 2007. Currently, he is an Associate Professor at Hokkaido University, Japan. His research interests include the area of robotics, navigation, sensor, and non-linear signal processing. He is a member of the IEEE, RSJ, JSME, and SICE.

Yukinori Kobayashi received his B.E., M.E., and PhD degrees in Mechanical Engineering from Hokkaido University, Japan, in 1981, 1983, and 1986, respectively. He is currently the President of the National Institute of Technology, Tomakomai College, Japan, and an Emeritus Professor of Hokkaido University, Japan. His research interests include vibration control of flexible structure, control problem of robots having flexibility, path planning, and navigation of mobile robots, vibration analysis, and non-linear vibrations of continuous systems. He is a member of the JSME, SICE, RSJ and EAJ.