CEKD: Cross-modal Edge-privileged Knowledge Distillation for Semantic Scene Understanding using Only Thermal Images

Zhen Feng, *Graduate Student Member, IEEE*, Yanning Guo, and Yuxiang Sun, *Member, IEEE*

***Abstract*—Semantic scene understanding using thermal images has received great attention due to the advantage that thermal imaging cameras could see in challenging illumination conditions. However, thermal images are lack of color information and the edges in thermal images are often blurred, making them not very suitable to be directly used by existing semantic segmentation networks that are designed with RGB images. To address this problem, we propose a cross-modal edge-privileged knowledge distillation framework, which utilizes a well-trained RGB-Thermal fusion-based semantic segmentation network with edge-privileged information as the teacher, to guide the training of a semantic segmentation network as the student. The student network only uses thermal images. The experimental results on a public dataset demonstrate that under the guidance of the teacher, the student network achieves superior performance over the state of the arts using only thermal images. Our code is available at https://github.com/lab-sun/CEKD.**

***Index Terms*—Knowledge Distillation, Semantic Segmentation, Privileged Information, Autonomous Driving, Thermal Images**

I. Introduction

As a fundamental technology in semantic scene understanding, semantic image segmentation has been widely studied in the field of autonomous driving [1]. The pixel-level classification results of semantic image segmentation serve as the basis for semantic scene understanding. The segmentation results can also be used for a variety of downstream tasks, such as path planning and trajectory prediction, etc.

With the advancement of deep-learning technologies, many deep-learning-based semantic segmentation networks have been proposed and have presented acceptable results [2]. These networks are mainly designed to use three-channel visible Red-Green-Blue (RGB) images [3]. However, the performance of existing RGB networks could be significantly degraded when the illumination conditions are unsatisfactory, such as nighttime, glare, on-coming headlights, etc. Using RGB and thermal image fusion could improve the performance of semantic segmentation under these challenging illumination conditions. So, many RGB-Thermal (RGB-T) fusion-based semantic segmentation networks [4–6] have been proposed in recent years.

However, multi-modal fusion requires precise sensor calibration, which could be affected during long-term use. For example, vehicle vibrations could change the extrinsic calibration parameters between RGB camera and thermal camera [7], hence leading to performance degradation for RGB-T fusion-based semantic segmentation [8]. Using only thermal images without sensor fusion could avoid this issue. In addition, most existing RGB-T fusion networks [4–6] use two independent encoders with similar structures, which increases the computation cost compared to those using only one modality of data. So, using only thermal images to achieve comparable performance to that of RGB-T fusion is a promising direction.

Although thermal imaging cameras could see in challenging illumination conditions, they still have shortcomings. For example, the object edges are often blurred [9] and the images are lacking of color information. These shortcomings make them not very suitable to be directly used by existing semantic segmentation networks that are designed with RGB images. Although some methods have been proposed to use only thermal images [9–13], the performance is still unsatisfactory and most of them ignore the fact that the pixel values in thermal images encode the temperature of objects.

To address the above issues, this letter proposes a cross-modal edge-privileged knowledge distillation framework that consists of an RGB-T fusion-based semantic segmentation network as the teacher, and a thermal-only semantic segmentation network as the student. Our motivation is to enable the thermal-only student network to achieve comparable performance to that of RGB-T fusion. This is realized by our knowledge distillation framework, which could transfer the edge detection capability from the teacher to the student. Note that the teacher is trained with edge detection ground truth as a kind of privileged information. Moreover, we believe that the pixel values in thermal images could be utilized to segment the background that usually has a temperature largely different from the foreground objects (e.g., cars, persons). So, we introduce a thermal enhancement (TE) module using the pixel values in the student network to increase the contrast between foreground and background to improve the segmentation performance. The main contributions of this letter are summarized as follows:

1) We propose a novel edge-privileged RGB-T fusion-based semantic segmentation network as the teacher.
2) We propose a cross-modal knowledge distillation method to train the thermal-only student network.

Manuscript received September 15, 2022; Revised December 13, 2022; Accepted January 29, 2023. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grants 62003286, 62273118, and 12150008, in part by the Hong Kong Innovation and Technology Fund under Grant ITS/145/21, in part by the Start-up Fund of The Hong Kong Polytechnic University under Grant P0034801. *(Corresponding author: Yuxiang Sun.)*

Zhen Feng is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, and the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, China (e-mail: zhen.feng@connect.polyu.hk).

Yanning Guo is with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, China (email: guoyn@hit.edu.cn).

Yuxiang Sun is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: yx.sun@polyu.edu.hk, sun.yuxiang@outlook.com).

Digital Object Identifier (DOI): see top of this page.

3) We propose a novel TE module in the student network to increase the contrast between foreground and background to improve the overall segmentation performance.

II. Related Work

A. Semantic Segmentation with RGB-T Fusion

He *et al.* [14] proposed MFNet in the two-encoders-one-decoder architecture and released an RGB-T fusion dataset for semantic segmentation. Sun *et al.* [15] proposed RTFNet using ResNet [16] as the encoder backbone. The features extracted from the thermal encoder are fused into RGB encoder via element-wise summation. Xu *et al.* [17] adopted an attention module in AFNet to fuse the feature maps extracted from RGB and thermal images. Zhang *et al.* [6] proposed ABMDRNet that fuses features of two modalities by reducing the difference between the two modalities.

B. Semantic Segmentation with only Thermal Images

The existing works on semantic segmentation using only thermal images mainly focus on how to overcome the problems of low resolution and lack of texture in thermal images. Li *et al.* [10] proposed EC-CNN that uses edge information to improve the segmentation accuracy. They employed HED [18] to generate the edge information. Ren *et al.* [9] also introduced edge information into MPSA, which fuses edge features into the output of the encoder and feeds the fusion results into the decoding module. Xiong *et al.* [11] proposed MCNet that combines attention modules and edge information. MCNet computed the loss of multi-level outputs for training. Munir *et al.* [12] proposed ARTSeg that combines attention modules and Recurrent Convolution Neural Network (RCNN). ARTSeg is designed in an encoder-decoder architecture. Each stage of the encoder and decoder is designed in the RCNN architecture. Kim *et al.* [13] proposed a multi-spectral unsupervised domain adaptation (MS-UDA) for thermal image semantic segmentation.

C. Edge Detection

Xie *et al.* [18] proposed HED that detects edges from multi-level features. They also proposed a class-balanced cross-entropy loss function to overcome the issue that the number of pixels of edges is much smaller than that of the background. Hu *et al.* [19] proposed a dynamic feature fusion strategy to detect the edge of objects. Su *et al.* [20] employed dilation convolutions and spatial attention modules to design PiDiNet for edge detection.

D. Knowledge Distillation

Knowledge distillation (KD) is often used to transfer knowledge from a large complex network (i.e., teacher network) to a small and simple network (i.e., student network) to achieve model compression [21]. Qin *et al.* [22] adopted the KD method to design a lightweight network with the region affinity distillation module for medical image segmentation. Wen *et al.* [23] adopted two KD modules to learn boundary information from a teacher network. KD is also used to learn knowledge across different modalities, such as transferring knowledge from the RGB modality to the thermal modalities.

III. The Proposed Network

A. The Overall Architecture

Fig. 1 shows the overall architecture of our proposed cross-modal edge-privileged knowledge distillation (CEKD) segmentation framework. As we can see, there are two sub-networks in our framework, namely, a cross-modal edge-privileged segmentation network (CENet) and an edge-privileged knowledge-distillation segmentation network (EKNet). CENet is a teacher network. It is an RGB-T fusion-based semantic segmentation network. EKNet is a student network. It is a thermal-only semantic segmentation network. EKNet learns the edge detection capability from the teacher network CENet. The encoder and decoder of EKNet have the same structure as CENet, except that the decoder of CENet is also fused with the predicted edge map $\hat{e}$. The $\hat{e}$ is predicted via an edge detection (ED) module. We refer readers to the paper [20] to get more details about the ED module. Edge label e is generated by an edge labels generation (ELG) module. The outputs of CENet are the predicted edge map $\hat{e}$ and the segmentation map $\hat{y}_{rt}$. The output of EKNet is the segmentation map $\hat{y}_t$ using only the thermal images.

B. The Teacher Network

The teacher network, CENet, is a cross-modal edge-privileged semantic segmentation network designed to fuse RGB images I_r and thermal images I_t. CENet consists of a five-stage RGB encoder, a five-stage thermal encoder, an ED module, and a five-stage decoder. The encoders and decoder are borrowed from RTFNet-50 [15]. We replace the last stage of the encoders with the last stage of BotNet [24]. The output of each stage of the thermal encoder is fused into the corresponding stage of the RGB encoder by element-wise addition. We denote the fusion result of the output of the n-th stage of the encoders as $f_{rt}^n(I_r, I_t), n \in 1, 2, \cdots, 5$, abbreviated as f_{rt}^n. We introduce skip connections between the RGB encoder and the decoder. We denote the output of the n-th stage of the decoder as $g_{rt}^n(g_{rt}^{n+1} \oplus f_{rt}^n), n \in 1, 2, 3, 4$, abbreviated as g_{rt}^n, where $\oplus$ represents an element-wise addition. Note that the output of the 5-th stage of the decoder is $g_{rt}^5(f_{rt}^5)$.

f_{rt}^5 is used to teach EKNet how to extract features. The $\hat{e}$ is predicted from $g_{rt}^3 \oplus f_{rt}^2$ via the ED module. The ED module is borrowed from [20]. The $\hat{e}$ is fused into the input of the ED module by element-wise addition. The fusion result is denoted as E_{rt}, which encodes the edge information. E_{rt} is fed into the second stage of the decoder and used to teach EKNet how to extract features with edge information. g_{rt}^1 is used to teach EKNet and generate the segmentation result $\hat{y}_{rt}$ through a Softmax layer. The edge labels are generated by the ELG module, which is borrowed from [19].

C. The Student Network

The student network, EKNet, is designed for semantic segmentation using only thermal images. EKNet also adopts a five-stage encoder and a five-stage decoder. The encoder and decoder share the same structure as the thermal encoder and decoder of CENet. There are also skip connections between the encoder and the decoder. We denote the output of the n-th stage of the encoders as $f_t^n(I_t), n \in 1, 2, \cdots, 5$, abbreviated as f_t^n. We denote the output of the n-th stage of the decoder

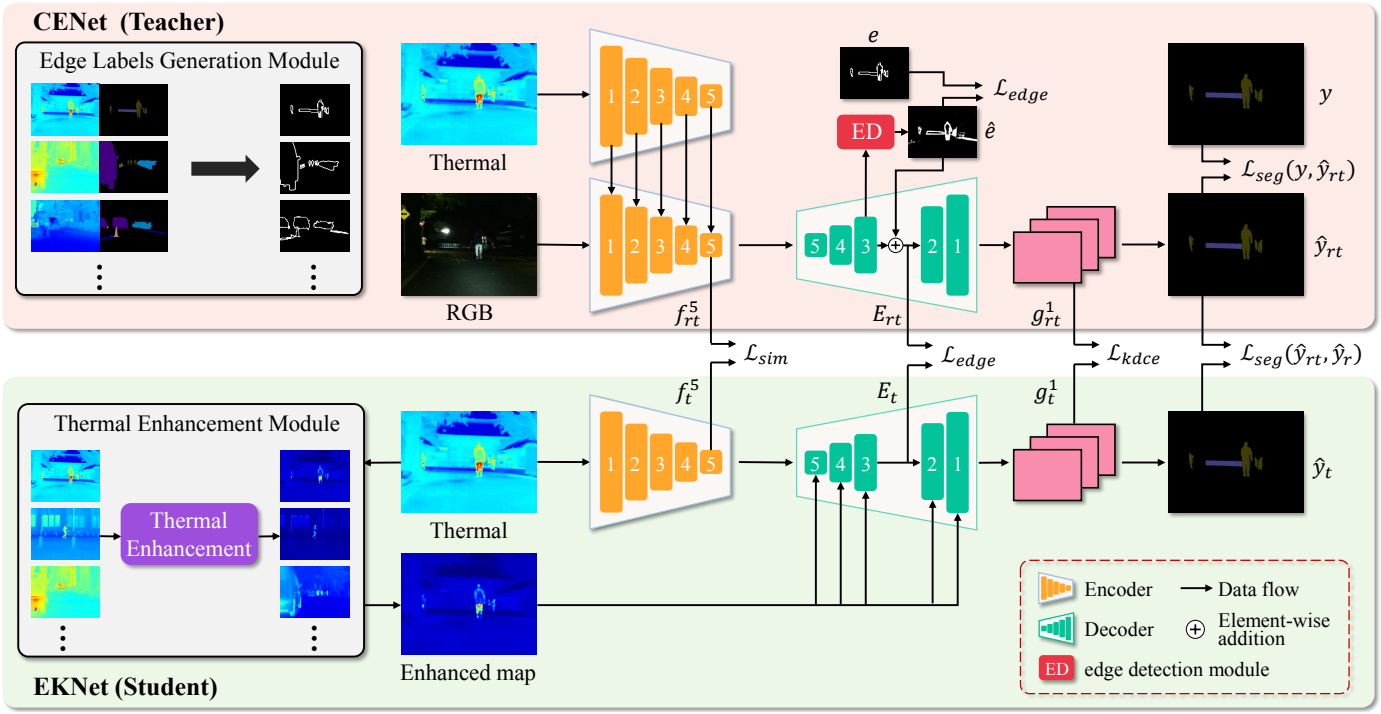

Fig. 1. The architecture overview of our cross-modal edge-privileged knowledge distillation framework. It consists of an RGB-T fusion network CENet (teacher network) and a thermal-only network EKNet (student network). The CENet consists of a five-stage RGB encoder, a five-stage thermal encoder, an edge detection (ED) module, and a five-stage decoder. The encoders and the decoder are borrowed from RTFNet [15]. We replace the last stage of both encoders with the last stage of BotNet [24]. The ED module is borrowed from [20]. Edge labels are generated by the ELG module borrowed from [19].

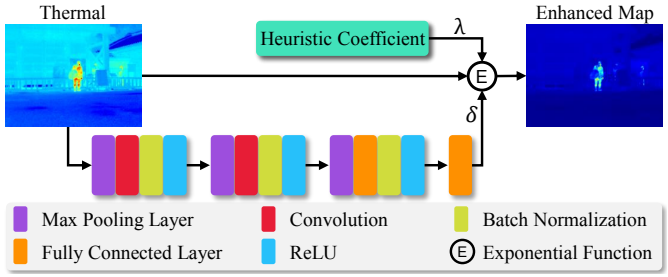

Fig. 2. The architecture of our proposed TE module. The core of the TE module is the exponential function, which increases the contrast between foreground (e.g., persons) and background in thermal images. The heuristic coefficient λ, the thermal image, and the exponent δ are fed into the exponential function.

as $g_t^n(g_t^{n+1} \oplus f_t^n(I_t)), n \in 1,2,3,4$, abbreviated as g_t^n. The encoder output f_t^5 is fed into the decoder. Note that the output of the 5-th stage of the decoder is $g_t^5(f_t^5(I_t))$. Thermal images are simultaneously fed into the encoder and the TE module. EKNet learns how to extract features from CENet, which means that f_t^5 and f_{rt}^5 should be the same. The enhanced map I_e generated by the TE module is downsampled into five different resolutions, i.e., $1/32$, $1/16$, $1/8$, $1/4$, and $1/2$ of the original resolution of the thermal images with the nearest interpolation method. The downsampled results are fused with the input of each stage of the decoder through element-wise addition. To learn the edge detection capability from the teacher network, the result of $g_t^3 \oplus f_t^2$, denoted as E_t, is needed to learn from E_{rt}. The output of the decoder g_t^1 and the segmentation result $\hat{y}_t$ also should learn knowledge from g_{rt}^1 and $\hat{y}_{rt}$, which means $\hat{y}_{rt}$ is pseudo labels for EKNet.

D. The TE Module

Thermal images encode the temperature information of the environment. The temperature of some objects to be segmented, such as *person*, is usually higher than that of the background. The TE module is proposed to increase the contrast between foreground (e.g., cars and persons) and background by an exponential function with an exponent greater than 1, which increases the contrast of the objects to the background. In the TE module, the pixel values of thermal images are normalized into $[0, 1]$, larger values indicate higher temperature. Since the pixel values are in the range of $[0, 1]$, all the pixel values still lie in the range of $[0, 1]$ after the exponential function, which means that all pixel values are still normalized. Smaller pixel values (e.g., trees) are closer to 0 (background) after the exponential function with an exponent greater than 1, which causes the objects (e.g., trees) to be blurred into the background. However, the contrast between larger pixel values (e.g., persons) to smaller pixel values is increased by the exponential function, which highlights the objects with higher temperatures (e.g., persons).

The structure of the TE module is shown in Fig. 2. The core of the TE module is an exponential function. Three data are fed into the exponential function, namely, the thermal image I_t, a heuristic coefficient λ, and the exponent δ. δ is adaptively generated from I_t. Firstly, I_t is fed into two consecutive identical blocks. Each block consists of a max pooling layer, a convolution layer, a batch normalization (BN) layer, and a ReLU layer. Secondly, the result is fed into a max pooling layer, a fully connected (FC) layer, a BN layer, and a ReLU layer. Finally, δ is generated by an FC layer. We denote the generation process as $\delta = G(I_t)$. λ ensures that the exponent of the exponential function is greater than 1 during the training and guides the generation of δ. The TE module is described as follows:

$$I_e(i,j) = \Big(I_t(i,j)\Big)^{\delta+\lambda}, \tag{1}$$

where $i = 1,2,3,\cdots,H$ and $j = 1,2,3,\cdots,W$. H and W represent the height and the width of I_t, respectively.

E. The Loss for CENet

In order to train CENet, we adopt the cross-entropy loss $\mathcal{L}_{seg}(y, \hat{y}_{rt})$ between the ground truth y and the segmentation result of CENet $\hat{y}_{rt}$, as well as the class-balanced cross-entropy loss [18] $\mathcal{L}_{edge}$ between the edge label e and the predicted edge $\hat{e}$. We combine $\mathcal{L}_{seg}(y, \hat{y}_{rt})$ and $\mathcal{L}_{edge}$ as the total loss $\mathcal{L}_t$ to train CENet, which is denoted as $\mathcal{L}_t = \mathcal{L}_{seg}(y, \hat{y}_{rt}) + \mathcal{L}_{edge}$, where $\mathcal{L}_{seg}(y, \hat{y}_{rt})$ and $\mathcal{L}_{edge}$ are calculated as follows:

$$\mathcal{L}_{seg}(y, \hat{y}_{rt}) = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W} y\big(i,j\big)\log\Big(\hat{y}_{rt}\big(i,j\big)\Big), \quad (2)$$

$$\mathcal{L}_{edge} = -\frac{1}{N}\Big(\alpha\sum_{i\in e_+}\log\left(\hat{e}_i\right) + \beta\sum_{i\in e_-}\log\left(1-\hat{e}_i\right)\Big), \quad (3)$$

where $N = H \times W$ represents the number of pixels of an image. $\alpha = e_-/e$, $\beta = e_+/e$, and $e = e_- + e_+$, where e_+ and e_- represent the pixel of the edge and background in the edge label e, respectively.

F. The Loss for Knowledge Distillation

We use a self-supervised manner to train EKNet. We use the result of CENet $\hat{y}_{rt}$ as pseudo labels. So, we adopt the cross-entropy loss $\mathcal{L}_{seg}(\hat{y}_{rt}, \hat{y}_t)$, which is computed in the same way as (2). To enable EKNet to learn more knowledge from CENet, we let EKNet learn the encoder output f_{rt}^5, edge fusion results E_{rt}, and decoder output g_{rt}^1 of CENet, which means learning the capabilities of feature extraction, edge detection, and object segmentation, respectively. We adopt the pair-wise similarity loss $\mathcal{L}_{sim}$ proposed in [25] for learning the feature extraction capabilities, the mean squared error (MSE) loss $\mathcal{L}_e$ for learning the detecting edge detection capability, and the knowledge-distillation cross-entropy loss $\mathcal{L}_{kdce}$ for learning the object segmentation capability. We combine these losses as the total knowledge distillation loss $\mathcal{L}_s$ to train EKNet, which is denoted as $\mathcal{L}_s = \mathcal{L}_{seg}(\hat{y}_{rt}, \hat{y}_t) + \mathcal{L}_{sim} + \mathcal{L}_e + \mathcal{L}_{kd}$. The function of the $\mathcal{L}_{sim}$ is denoted as follows:

$$\mathcal{L}_{sim} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_{ij}^{rt} - a_{ij}^{t}\right)^2, \quad (4)$$

where a_{ij}^{rt} (a_{ij}^{t}) represents the similarity between the i-th pixel and the j-th pixel produced from f_{rt}^5 (f_t^5). The similarity between two pixels $\mathbf{f}_i$ and $\mathbf{f}_j$ is calculated as $a_{ij} = \frac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_2\|\mathbf{f}_j\|_2}$. We hope that the last three stages of the decoder of the student network can learn the capability of edge detection and information recovery from the teacher network, that is, let E_t be the same as E_{rt}, so we use loss $\mathcal{L}_e$ to guide E_t to be close to E_{rt}. The $\mathcal{L}_e$ is calculated as follows:

$$\mathcal{L}_e = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(E_{rt}(i,j) - E_t(i,j)\right)^2, \quad (5)$$

where E_{rt} and E_t have the same shape, that is, the numbers of channel, height, and width are 256, 120, and 160. The loss $\mathcal{L}_{kdce}$ allows the knowledge of a teacher network to be distilled into a student network at high temperatures. The $\mathcal{L}_{kdce}$ is calculated as follows:

$$\mathcal{L}_{kdce} = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\sigma\Big(\frac{g_{rt}^1(i,j)}{T}\Big)\log\left(\sigma\Big(\frac{g_t^1(i,j)}{T}\Big)\right), \quad (6)$$

where T represents the temperature of distillation, $\sigma(\cdot)$ represents the Softmax function.

TABLE I

THE RESULTS (%) OF THE ABLATION STUDY ON THE ED MODULE. *Input* INDICATES WHERE THE INPUT OF THE ED MODULE COMES FROM. *Output* INDICATES WHERE THE OUTPUT OF THE ED MODULE GOES INTO. '–' MEANS THAT NO DATA IS COMING FROM THIS STRUCTURE OR FUSED WITH THE OUTPUT OF THIS STRUCTURE.

Variants	Input		Output	IoU_e	mAcc	mIoU
	Encoder	Decoder	Decoder			
NED	–	–	–	0.0	64.6	54.0
E123D3	1, 2, 3	–	3	20.4	69.1	55.1
E123D2	1, 2, 3	–	2	11.4	68.5	54.8
D234D2	–	2, 3, 4	2	17.0	71.2	55.0
E2D3	2	–	3	16.0	65.1	54.1
E1D2	1	–	2	6.6	68.4	54.4
D2D2	–	2	2	19.4	69.6	55.3
D3D3 (Ours)	–	3	3	**35.13**	**71.8**	**56.1**

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. The Dataset

We use the public MFNet dataset released in [14] for experiments. The MFNet dataset has 9 classes (including the unlabelled background class) of manually annotated labels for semantic segmentation. It contains $1,596$ pairs of RGB and thermal images, among which 820 pairs are captured at the daytime and 749 pairs are captured at nighttime. We adopt the same split scheme in [14] to train our network. We adopt the method proposed in [19] to generate edge labels with width 1 based on the manually-labeled semantic masks. To avoid the further downsampling on edge labels during training. We generate the edge labels on the downsampled semantic masks ($1/4$ and $1/2$ of the original resolution).

B. Training Details

We implement our proposed method using PyTorch. The networks are trained with NVIDIA RTX 3090. We borrow the initialization scheme of RTFNet to initialize both the networks except the last stages of encoders, ED module, and TE module. Other parameters are initialized by the default scheme of PyTorch. We adopt the same set of parameters as RTFNet for the optimization solver, such as initial learning rate, decay strategy, etc. The batch size is set to 2 during training.

Firstly, we train the teacher network separately to obtain the best-performing teacher network. Note that we first train CENet using the loss $\mathcal{L}_{seg}(y, \hat{y}_{rt})$ without the ED module as a baseline. Then, we adopt the loss $\mathcal{L}_t$ and $\mathcal{L}_{seg}(y, \hat{y}_{rt})$ to train CENet with the pre-trained weight of the baseline. Secondly, we use the teacher network to guide the training of the student network (i.e., knowledge distillation). During the process of knowledge distillation, the parameters of the student network are initialized by the aforementioned schemes.

During the training, we fix the parameters of the teacher network and only optimize the parameters of the student network. We adopt the same metrics, Accuracy (Acc) and Intersection-over-Union (IoU), from [15] to evaluate our network.

C. Ablation Study

1) Ablation on the ED Module: For the ED module, we first remove the ED module from CENet to demonstrate the benefits brought by the ED module. We name this variant as NED (no edge detection). In addition, we change the position of the ED module in the network, that is, we change the position where the input data of the ED module comes from and the

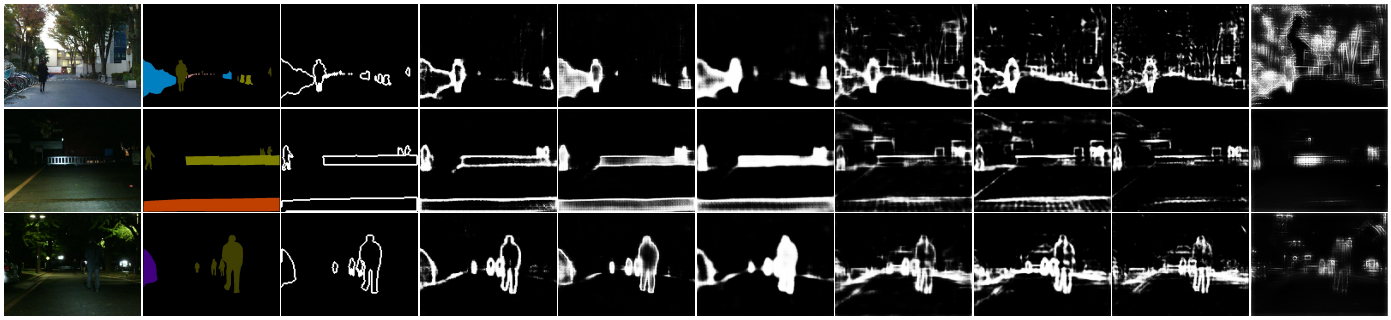

Fig. 3. Sample qualitative demonstrations of the edge detection results. GT represents ground truth labels. Each row represents the edge detection results from different variants with the same input image. The results show that our D3D3 (CENet) can segment the edges of objects with better performance.

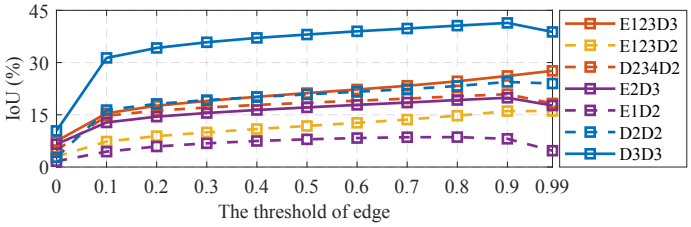

Fig. 4. The results (%) of IoU for edge detection with different thresholds.

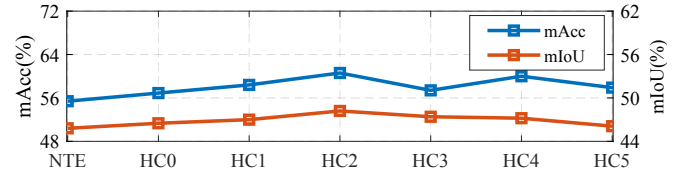

Fig. 5. The results (%) of mAcc and mIoU for variants with different values of the heuristic coefficient. Variants are trained via a supervised training method. *NTE* means that there is no TE module in the variant. *HC1* means that the value of the heuristic coefficient is 1.

position where the output data is fed into the network. We design variants according to the following two rules: 1) We use edge information to guide the decoder to produce segmentation masks, so the edge information should be only fused with the decoder (g_{rt}^2, g_{rt}^3, g_{rt}^4, and g_{rt}^5), and should not be fused with the decoder output g_{rt}^1; 2) The edge information from the small-size feature maps could be weak, so the edge information should not be extracted from the small-size feature maps (f_{rt}^5, f_{rt}^4, and g_{rt}^5). Moreover, the edge information should not be fused with small-size feature maps (g_{rt}^5 and g_{rt}^4), because the edge information is weak. The details of the variants are listed as follows:

1) *E123D3*: The inputs of the ED module are f_{rt}^1, f_{rt}^2, and f_{rt}^3. The output of the ED module is fused with $g_{rt}^3 \oplus f_{rt}^2$.
2) *E123D2*: The inputs of the ED module are f_{rt}^1, f_{rt}^2, and f_{rt}^3. The output of the ED module is fused with $g_{rt}^2 \oplus f_{rt}^1$.
3) *D234D2*: The inputs of the ED module are $g_{rt}^2 \oplus f_{rt}^1$, $g_{rt}^3 \oplus f_{rt}^2$, and $g_{rt}^4 \oplus f_{rt}^3$. The output of the ED module is fused with $g_{rt}^2 \oplus f_{rt}^1$.
4) *E2D3*: The input of the ED module is f_{rt}^2. The output of the ED module is fused with $g_{rt}^3 \oplus f_{rt}^2$.
5) *E1D2*: The input of the ED module is f_{rt}^1. The output of the ED module is fused with $g_{rt}^2 \oplus f_{rt}^1$.
6) *D2D2*: The input of the ED module is $g_{rt}^2 \oplus f_{rt}^1$. The output of the ED module is fused with $g_{rt}^2 \oplus f_{rt}^1$.

Following the naming rule, our CENet can also be called *D3D3*. All the variants and our CENet are trained and tested with the MFNet dataset and the generated edge labels. Tab. I

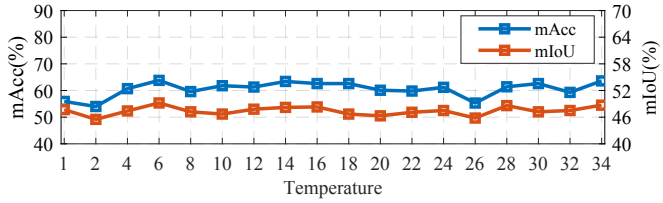

Fig. 6. The results (%) of mAcc and mIoU of EKNet trained at different temperatures in the knowledge distillation process. We can see that our EKNet achieves the best performance when the distillation temperature is 6.

displays the results of all the variants and our CENet. Compared with *NED* and other variants, we can see that the introduction of the ED module improves the performance of the networks. Our CENet gets the best results compared with all the variants. In addition, when the input data of the ED module comes from the decoder, the results are generally better than the case when the input data come from the encoder. The reason may be that features in the encoder contain background features, and the edges detected from the encoder provide the background information into the decoder. We display the qualitative edge detection results in Fig. 3. Each row of Fig. 3 is the edge detection results for different variants of the same image. From Fig. 3, the results of *E123D2*, *E123D2*, *E2D3*, and *E1D2* contain the edge of the background. However, the results of *D3D3*, *D2D2*, and *D234D2* mainly contain the edge of objects. The qualitative results show that the edges detected from the encoder feature map contain more background edges.

We also calculate the IoU of edges to evaluate the accuracy of edge detection. The values of the predicted edge map of the ED module are normalized to $[0, 1]$. We generate edges in the predicted edge map of the ED module for all the variants by thresholding the feature maps with threshold values 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99, respectively. When the pixel value of the output is greater than the threshold, the pixel is considered as a part of an edge. The results of IoU for edge detection with different thresholds are displayed in Fig. 4. We use the mean of these results (IoU_e) as a metric to evaluate the performance. From Tab. I, we can find that our D3D3 achieves the best results with different thresholds. Overall, *D3D3* (i.e., our CENet) can accurately detect the edge of objects.

2) Ablation on the TE module: For the TE module, we tune the heuristic coefficient. It is one of the exponents of the exponential function in the TE module. Large exponents would cause more pixels in the thermal image to approach 0 with the exponential function. Therefore, we try the value of the heuristic coefficient from 0 to 6. We use *HC1* to represent the variant with

TABLE II
THE RESULTS (%) OF THE ABLATION ON LOSSES. – AND ✓MEAN THE LOSS IS NOT USED FOR TRAINING AND IS USED FOR TRAINING, RESPECTIVELY.

Experiment	$\mathcal{L}_{seg}$	$\mathcal{L}_{kdce}$	$\mathcal{L}_{sim}$	$\mathcal{L}_{e}$	mAcc	mIoU
(A)	✓	–	–	–	59.5	46.5
(B)	✓	✓	–	–	62.4	48.1
(C)	✓	–	✓	–	60.2	47.1
(D)	✓	–	–	✓	61.3	47.9
(E)	✓	✓	✓	–	63.0	48.5
(F)	✓	✓	–	✓	63.5	48.8
(G)	✓	✓	$\mathcal{L}_{afd}$	✓	62.3	49.1
(H)	✓	✓	$\mathcal{L}_{pad}$	✓	63.3	48.4
(I)	✓	✓	✓	✓	**63.8**	**49.2**

TABLE III
THE RESULTS (%) OF THE ABLATION ON TRAINING STRATEGIES. ✓AND – MEAN THAT THE TEACHER NETWORK CONTAINS AND NOT CONTAINS THE MODAL, RESPECTIVELY.

Strategy	Supervision method	Teacher		mAcc	mIoU
		RGB	Thermal		
(A)	Supervised	–	–	60.6	48.2
(B)	Self-Supervised	–	✓	57.5	46.4
(C)	Self-Supervised	✓	✓	**63.8**	**49.2**

the heuristic coefficient being 1. We also remove the TE module from EKNet to demonstrate the benefits of the TE module, which is denoted as *NTE*. Fig. 5 shows the results of the ablation study. The mAcc and mIou of *NTE* are the smallest among all the variants, indicating that the TE module could improve the performance of the networks. We can also see that *HC2* achieves the best results. So, we set the heuristic coefficient of the TE module in EKNet to 2.

3) Ablation on the Temperature of Knowledge Distillation: We distill the knowledge of CENet to EKNet at different temperatures to enable EKNet to achieve the best performance. Besides setting the distillation temperature to 1, we try even numbers from 2 to 34. Fig. 6 shows the results of EKNet distilled at different temperatures. The results show that the appropriate temperature of knowledge distillation can enable the student network to learn more knowledge from the teacher network. From the results, it can be found that EKNet achieves the best performance when the distillation temperature is 6.

4) Ablation on Losses: We conduct ablation studies on losses to verify the effectiveness of different combinations of losses. There are 4 losses during the training process of knowledge distillation, namely $\mathcal{L}_{seg}$, $\mathcal{L}_{kdce}$, $\mathcal{L}_{sim}$, and $\mathcal{L}_{e}$. We train the student network with different combinations of losses. Moreover, we replace $\mathcal{L}_{sim}$ respectively with the loss $\mathcal{L}_{afd}$ [26] and the loss $\mathcal{L}_{pad}$ [27] to train the student network. The details and results of each experiment are presented in Tab. II. we can see that the student network cannot achieve optimal performance with only one loss. Comparing the results of (E) with (I), we can find that the student network learns the edge detection capability through the loss $\mathcal{L}_{e}$, which improves the performance of the network. Comparing the results of (G), (H), and (I), we can find that the loss $\mathcal{L}_{sim}$ enables the student network to achieve better performance.

5) Ablation on Training Strategies: In the ablation study, we use three strategies to train the student network (EKNet): 1) Train EKNet with a supervised method using only thermal images; 2) Train EKNet with a self-supervised method using a thermal-only teacher network; 3) Train EKNet with a self-supervised method using an RGB-Thermal fusion teacher network (CENet). We remove the RGB encoder from CENet as the thermal-only teacher network. The training details and results are displayed in Tab. III.

TABLE IV
THE COMPARATIVE RESULTS (%) OF THE MULTI-MODAL NETWORKS ON DAYTIME AND NIGHTTIME SCENARIOS.

Methods	Daytime		Nighttime	
	mAcc	mIoU	mAcc	mIoU
UNet++ [28]	56.9	**48.4**	57.2	51.0
MFNet [14]	42.5	38.0	42.9	38.6
FuseNet [29]	47.3	41.7	43.3	40.4
RTFNet [15]	57.3	44.4	59.4	52.0
SegHRNet [30]	42.1	37.0	41.8	38.3
CENet (Ours)	**61.5**	48.0	**68.8**	**56.1**

TABLE V
THE COMPARATIVE RESULTS (%) OF THE THERMAL-ONLY NETWORKS ON DAYTIME AND NIGHTTIME SCENARIOS.

Methods	Daytime		Nighttime	
	mAcc	mIoU	mAcc	mIoU
UNet++ [28]	49.5	39.4	58.5	47.4
MFNet [14]	33.9	26.9	32.6	30.0
FuseNet [29]	34.2	27.8	33.9	31.8
RTFNet [15]	47.4	37.9	50.8	44.9
SegHRNet [30]	38.6	31.2	44.0	40.7
MCNet [11]	35.6	29.7	38.6	35.7
EKNet (Ours)	**51.5**	**40.0**	**62.3**	**50.6**

Comparing (A) with (C), we can find that the network trained with the self-supervised knowledge distillation with the RGB-Thermal fusion teacher network achieves better results than that trained with the supervised method using thermal-only images. This demonstrates that EKNet effectively learns knowledge from the RGB-thermal fusion teacher network with knowledge distillation, such as the capability of edge detection in CENet that could not be learned directly from thermal images. The results of (B) show that when the teacher network lacks RGB information, it cannot guide the student network well. This again demonstrates that (C) learns RGB-thermal fused knowledge from the RGB-thermal fusion teacher network via distillation, which could not be learned from thermal images.

D. Comparative Study

We compare our proposed CENet and EKNet with the well-known networks, including UNet++ [28], MFNet [14], FuseNet [29], RTFNet [15], SegHRNet [30], ABMDRNet [6], AFNet [17], and MCNet [11]. We train UNet++ and SegHRNet with the 4-channel RGB-T images, so that they can be compared with other multi-modal networks. We train all the multi-modal networks with the RGB encoders removed to compare with our EKNet. MCNet is only trained for comparison with our EKNet.

1) The Overall Results for RGB-T Segmentation: Tab. VI displays the results of all the networks trained and tested with the RGB-T images with the MFNet dataset [14]. We treat the unlabeled background as one class and count it into the mAcc and mIoU calculation. Note that the results for the background are not displayed in Tab. VI, because they are all around 95 or higher. We directly import the results of ABMDRNet and AFNet from the original papers, as they are not open-sourced. From the results, we can see that our proposed CENet achieves the best performance in terms of both mAcc and mIoU, ABMDRNet

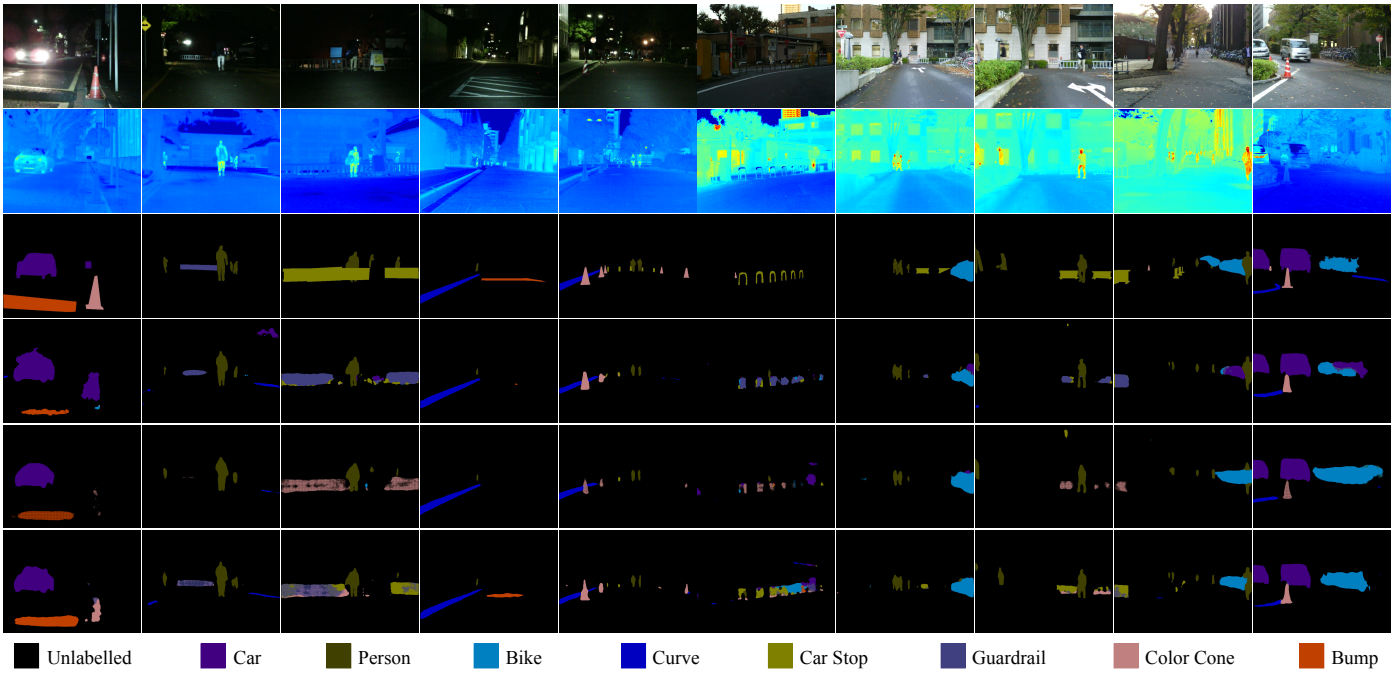

Fig. 7. Sample qualitative demonstrations. The rows from top to bottom are RGB images, thermal images, ground truth, UNet++ results, RTFNet results, and our EKNet results. RGB images are only used for display and not for training. The thermal images are colored with the *jet* color map.

TABLE VI
THE COMPARATIVE RESULTS OF RGB-T FUSION NETWORKS WITH MFNET DATASET [14]. UNET++ AND SEGHRNET IS TRAINED WITH 4-CHANNEL DATA. * REPRESENTS THAT THE RESULTS ARE IMPORTED FROM THE PAPERS [6] AND [17]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT.

Methods	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
UNet++ [28]	90.5	85.4	76.6	69.7	**75.8**	**62.1**	56.2	43.6	31.4	26.2	18.9	3.2	45.8	40.2	51.6	47.0	60.7	52.8
MFNet [14]	76.4	68.0	61.5	56.5	61.8	49.9	37.1	32.4	13.4	11.0	0.0	0.0	37.7	33.5	33.2	31.8	46.7	42.2
FuseNet [29]	76.3	74.4	72.6	66.5	68.9	57.1	44.3	39.0	14.5	13.1	0.0	0.0	41.6	31.8	7.6	7.5	47.3	43.0
RTFNet [15]	91.3	**86.3**	78.2	67.8	71.5	58.2	59.8	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	**57.2**	62.2	51.7
SegHRNet [30]	79.6	74.3	64.0	58.6	50.3	39.8	45.0	36.4	18.7	16.2	0.0	0.0	29.5	25.2	28.2	27.6	46.1	41.7
ABMDRNet* [6]	**94.3**	84.8	**90.0**	69.6	75.7	60.3	64.0	45.1	**44.1**	**33.1**	31.0	5.1	**61.7**	47.4	66.2	50.0	69.5	54.8
AFNet* [17]	91.2	86.0	76.3	67.4	72.8	62.0	49.8	43.0	35.3	28.9	24.5	4.6	50.1	44.9	61.0	56.6	62.2	54.6
CENet (Ours)	92.0	85.8	78.9	**70.0**	74.9	61.4	**64.8**	**46.8**	39.8	29.3	**65.7**	**8.7**	54.1	**47.8**	**77.1**	56.9	**71.8**	**56.1**

achieves the second best performance, and AFNet ranks the third place. Overall, our proposed CENet improves mAcc by 2.3% and mIoU by 1.3%.

2) The Overall Results for Thermal-only Segmentation: Tab. VII displays the results of all the networks trained and tested with thermal images only. We remove the RGB encoder of the multi-modal networks and change the input channel of the first layer of these networks to 1. We use the last output of MCNet to train itself. We also design one more experiment: we use the knowledge distillation method to train EKNet without the TE module, which is named KDNet.

Comparing the results of KDNet and EKNet, we can see that the network with the TE module achieves better results with the knowledge distillation method. EKNet achieves the best results in the *person* class, which also shows that the TE module provides benefits for EKNet to segment the *person* by highlighting the pixels of the class *person*. Comparing the best results of well-known networks, our proposed EKNet improves mAcc by 4.3% and mIoU by 2.3%.

3) The Daytime and Nighttime Results: We also evaluate the multi-modal networks and thermal-image networks with daytime and nighttime images. Tab. IV displays the results of the multi-modal networks, and Tab. V displays the results of the thermal-only networks. From Tab. IV, although the mIoU for the daytime of our network is slightly lower than UNet++, other metrics are much higher than UNet++, which are the best results. Tab. V shows that our EKNet achieves the optimal results in both scenarios.

4) The Qualitative Demonstrations for Thermal-only Segmentation: Fig. 7 qualitatively demonstrate the three networks with the best results trained with only thermal images. In Fig. 7, the first 5 columns are images at nighttime, and the last 5 columns are images at daytime. The thermal image in the column 6 is so challenging that all networks incorrectly segment the class *color cone* and *color stop*. In contrast, our EKNet achieves the best result. We can see that a part of class *color stop* is correctly segmented. The column 4 shows that our EKNet is the only network that can segment class *bump*. Our network is also the only one that can segment class *color cone* in the first column. From the results, we could find that class *color stop*, *color cone*, and *bump* are big challenges. However, our EKNet still achieves better results.

V. CONCLUSIONS AND FUTURE WORK

We presented here a cross-modal edge-privileged knowledge distillation framework, which transfers the edge detection capa-

TABLE VII

THE COMPARATIVE RESULTS OF NETWORKS TRAINED WITH ONLY THERMAL IMAGES FROM THE MFNET DATASET [14]. ALL THE NETWORKS USE GROUND-TRUTH LABELS THROUGH SUPERVISED LEARNING, EXCEPT KDNET AND EKNET. KDNET AND EKNET ARE TRAINED USING THE SAME TEACHER NETWORK AND THE SAME TRAINING METHOD. KDNET IS A VARIANT OF EKNET WITHOUT THE TE MODULE.

Methods	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
UNet++ [28]	**84.5**	76.5	74.8	66.5	55.5	48.2	48.2	36.9	14.1	11.5	**47.9**	3.5	44.6	**30.2**	66.9	51.2	59.5	46.9
MFNet [14]	52.7	46.1	53.1	48.4	41.1	29.4	34.4	30.4	4.1	3.8	0.0	0.0	6.9	6.1	18.6	14.3	34.4	30.4
FuseNet [29]	28.0	27.8	48.3	45.7	28.8	27.3	29.0	26.7	6.2	6.1	0.0	0.0	26.6	23.3	53.9	40.1	35.6	32.4
RTFNet [15]	83.3	78.4	74.1	64.2	**64.6**	50.0	**52.4**	**40.5**	13.5	11.5	0.0	0.0	25.4	11.9	54.2	44.0	51.8	44.2
SegHRNet [30]	66.3	61.4	66.9	59.3	30.6	28.7	42.3	35.1	6.6	5.9	1.6	0.3	21.6	15.4	61.6	46.9	44.1	38.8
MCNet [11]	57.6	51.6	62.0	53.8	26.6	21.7	30.1	26.8	1.4	1.3	6.0	2.1	12.1	6.7	49.1	46.7	38.2	34.0
KDNet (Variant)	83.3	78.4	75.0	65.6	64.2	50.0	51.4	39.6	27.0	19.7	31.5	3.7	41.3	27.3	77.8	54.5	61.2	48.5
EKNet (Ours)	82.7	**78.6**	**78.6**	**67.5**	63.6	**51.9**	51.1	39.3	**28.3**	**20.1**	47.7	**7.1**	**45.1**	25.2	**78.3**	**55.9**	**63.8**	**49.2**

bility of an RGB-T teacher network to a thermal-only student network. Specifically, we first proposed a novel edge-privileged RGB-T fusion network as the teacher by introducing an edge detection module into the decoder. The edge detection ground truth is used as the privileged information during the training of the teacher. Secondly, we proposed a novel thermal-only semantic segmentation network with our proposed TE module as the student. Finally, we used a knowledge distillation method to distill the edge detection capability of the teacher network to the student network to achieve our goal.

However, our method still has some limitations. First, we use an exponential function with the same parameters for all the pixels in the TE module, making it difficult to highlight objects that have a small temperature difference from the background. We would like to design power exponents for each sub-regions in the future to alleviate this issue. Second, we only use the output of the TE module in the decoder, so the student network may not well utilize the thermal enhancement information. We would like to introduce the information into both the encoder and decoder in the future.

REFERENCES

[1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[2] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[3] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[4] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[5] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for rgb-t semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[6] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *CVPR*, 2021, pp. 2633–2642.

[7] R. Usamentiaga and D. Garcia, "Multi-camera calibration for accurate geometric measurements in industrial environments," *Measurement*, vol. 134, pp. 345–358, 2019.

[8] A. Kundu, X. Yin, A. Fathi, D. Ross, B. Brewington, T. Funkhouser, and C. Pantofaru, "Virtual multi-view fusion for 3d semantic segmentation," in *ECCV*, 2020, pp. 518–535.

[9] S. Ren, Q. Liu, and X. Zhang, "Mpsa: A multi-level pixel spatial attention network for thermal image segmentation based on deeplabv3+ architecture," *Infrared Physics & Technology*, vol. 123, p. 104193, 2022.

[10] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, 2020.

[11] H. Xiong, W. Cai, and Q. Liu, "Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infrared Physics & Technology*, vol. 113, p. 103628, 2021.

[12] F. Munir, S. Azam, U. Fatima, and M. Jeon, "Artseg: Employing attention for thermal images semantic segmentation," in *Asian Conference on Pattern Recognition*, 2022, pp. 366–378.

[13] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon, "Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6497–6504, 2021.

[14] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IROS*, 2017, pp. 5108–5115.

[15] Y. Sun, W. Zuo, 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.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] J. Xu, K. Lu, and H. Wang, "Attention fusion network for multi-spectral semantic segmentation," *Pattern Recognition Letters*, vol. 146, pp. 179–184, 2021.

[18] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015, pp. 1395–1403.

[19] Y. Hu, Y. Chen, X. Li, and J. Feng, "Dynamic feature fusion for semantic edge detection," *arXiv preprint arXiv:1902.09104*, 2019.

[20] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu, "Pixel difference networks for efficient edge detection," in *ICCV*, 2021, pp. 5117–5127.

[21] B. Li, S. Wang, H. Ye, X. Gong, and Z. Xiang, "Cross-modal knowledge distillation for depth privileged monocular visual odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6171–6178, 2022.

[22] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[23] Y. Wen, L. Chen, S. Xi, Y. Deng, X. Tang, and C. Zhou, "Towards efficient medical image segmentation via boundary-guided knowledge distillation," in *International Conference on Multimedia and Expo.*, 2021, pp. 1–6.

[24] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *CVPR*, 2021, pp. 16 519–16 529.

[25] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *CVPR*, 2019.

[26] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *AAAI*, vol. 35, no. 9, 2021, pp. 7945–7952.

[27] Y. Zhang, Z. Lan, Y. Dai, F. Zeng, Y. Bai, J. Chang, and Y. Wei, "Prime-aware adaptive distillation," in *ECCV*. Springer, 2020, pp. 658–674.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[29] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*, 2016, pp. 213–228.

[30] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv:1904.04514*, 2019.