# Capitalizing on RGB-FIR Hybrid Imaging for Road Detection

Yigong Zhang, Jin Xie, José M. Álvarez, Cheng-Zhong Xu, *Fellow, IEEE*, Jian Yang, and Hui Kong

***Abstract*— Traditionally, road detection approaches mostly capitalize on RGB images, 3D LiDAR point cloud or their fusion. However, RGB camera is sensitive to light conditions, while LiDAR point cloud is sparse compared with dense image pixels. In this work, a new hybrid image dataset is provided for the task of road detection based on cameras. In this dataset, the hybrid images are acquired by an optically aligned hybrid imaging device, consisting of a far-infrared (FIR) imager and an RGB camera to output pixel-wise registration of thermal and RGB frames. Then we investigate on three methods based on fully convolutional neural network (F-CNN) to demonstrate the advantages by fusing RGB-FIR images in road detection. First, a middle-fusion based model is built, where the output feature maps of encoder branches from RGB and FIR images are directly concatenated into a single-fusion branch as the decoder. Next, the originally discarded layers after fusion operation for both RGB and FIR branches are recovered as the mimic branches to imitate the distributions of the fusion outputs, which constitutes an extended cross model (ECM). Moreover, the outputs of mimic branches at different scales are also used to imitate the corresponding outputs in the fusion branch, called a hierarchical cross model (HCM). The experimental results demonstrate the effectiveness and efficiency of our fusion strategies.**

***Index Terms*— RGB-FIR fusion, road detection, CNN, extended cross model, hierarchical cross model.**

## I. Introduction

With the rapid development of sensor technology and computer vision, autonomous driving technique has become a research hot-shot in recent years. For an autonomous vehicle, one of the most critical task is road detection. It provides free space for planning motion and navigation. In spite of various achievements based on different types of sensors, there still exist some problems for accurate and robust detection, such as the ones caused by illumination conditions and significant changes in driving scenarios.

RGB camera is one of the most widely-used sensors for road detection. It outputs images with rich color and texture information in high resolution. However, RGB camera is very susceptible to illumination variation. Night, over-exposure or shadows all result in poor imagery, which is hard to be applied to robust road detection. Another widely used sensor for road detection is Light Detection and Ranging (LiDAR). It can emit multiple rays of laser light to obtain accurate 3D structure of the scene in a form of 3D point cloud, but it is sparse and unordered for users. The high expense of multi-ray LiDAR sensor also restricts its wide usage on the existing commercial.

Far infrared (FIR) or thermal camera is also a kind of popular sensor. It captures temperature information of the scene, achieving insensitivity to illumination variation. Compared with multiple-ray LiDAR sensors, the cost of FIR camera is cheaper, and the output image is as dense as that of RGB camera. Some exemplar images are shown in Figure 1. No matter during the day or night, or in the sunlight or shadow, FIR camera could provide clear images. Nevertheless, unlike the RGB images taken at daytime exhibiting rich colors and abundant textures, FIR images lack color and rich texture information. Little difference in temperature or infrared radiation may blur the boundaries of objects, such as the brick sidewalk in the first column of Figure 1(a), the white motorcycle body and the white guardrails in the last two columns of Figure 1(b). Intuitively, it is wise to use both RGB and FIR images for road detection because they can offer complementary information to improve performance for road scene perception.

Owing to the recent progress in developing coaxial camera system [1], [2], it is much easier to acquire pixel-wisely aligned RGB and FIR images. Whereas, most of the current studies focused on pedestrian detection [3]–[5], re-identification [6]–[8], face recognition [9], [10] and tracking [11], [12]. As far as we know, there are very few research works on road detection yet.

On the other hand, most of the F-CNN based road detection approaches generally merge the feature maps from different sensors into a single fusion branch to provide the final results, in which the original output branches for each sensor are omitted. Nevertheless, we find that some regions may be recognized correctly by a single sensor while be wrongly

Manuscript received 13 January 2021; revised 18 August 2021; accepted 20 October 2021. Date of publication 30 November 2021; date of current version 9 August 2022. The work of Jian Yang was supported in part by the National Natural Science Foundation of China under Grant U1713208 and in part by the Program for Changjiang Scholars. The Associate Editor for this article was N. Zheng. *(Corresponding author: Jian Yang.)*

Yigong Zhang, Jin Xie, and Jian Yang are with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and the Jiangsu Key Laboratory of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zyg025@hotmail.com; csjxie@njust.edu.cn; csjyang@njust.edu.cn).

José M. Álvarez is with NVIDIA Corporation, Santa Clara, CA 95051 USA (e-mail: jalvarez.research@gmail.com).

Cheng-Zhong Xu is with the State Key Laboratory of Internet of Things for Smart City (SKL-IOTSC), Department of Computer and Information Science, University of Macau, Macau (e-mail: czxu@um.edu.mo).

Hui Kong is with the State Key Laboratory of Internet of Things for Smart City (SKL-IOTSC), Department of Electromechanical Engineering (EME), University of Macau, Macau (e-mail: huikong@um.edu.mo).

Digital Object Identifier 10.1109/TITS.2021.3129692

Fig. 1. Some exemplar images in our RGB-FIR road dataset captured during (a) the daytime and (b) the night. For each sub-figure, the top row shows RGB images, the ones in the middle row are FIR images, and the ones in the bottom row are ground-truth images.

classified by sensor fusion. That means, some unique useful information can be extracted from a single specific sensor for road detection, but not learned by fusion during the learning process. Moreover, due to the various sceneries, it is impossible and unreasonable to directly choose which sensor is the most proper for a specific road scene.

To address these issues, we concentrate on utilizing RGB-FIR information for road detection. An RGB-FIR hybrid image dataset with ground-truth labeling is provided for road detection task, which is pixel-wisely aligned and attenuates the ghost phenomena caused by camera's rolling shutter. To our best knowledge, this is the first RGB-FIR dataset for road detection task. Then we investigate on three fusion strategies. First, we construct a simple concatenation based fusion model, which directly concatenates the output feature maps from the RGB and FIR branches into a single fusion branch. Inspired by the work [13], we restore the layers of RGB and FIR branches that has been removed after fusion operation as the mimic branches to imitate the distributions of the fusion outputs, which constitutes an extended cross model (ECM). By balancing the cross model difference and the segmentation loss, the network obtains the private information from each individual sensor can also provide better results. Moreover, because the decoder of an image based semantic segmentation network has feature maps at different scales, the outputs of mimic branches of different scales are also used to imitate the corresponding outputs in the fusion branch, called a hierarchical cross model (HCM).

Our contributions are listed as follows:

- We build a new RGB-FIR image dataset for road detection. The dataset consists of various types of road scenes under different illumination conditions. We also provide a strategy to reduce the ghost phenomenon.
- We design two network structures, the ECM and HCM for RGB-FIR fusion. These make the network get more specific and unique information from each individual sensor.
- On our road dataset, we validate several concatenation-based fusion models which are concatenated at different stages. The results indicate that the middle-fusion based method could acquire the best performance.
- We compare our middle-fusion based ECM and HCM models with the state-of-the-art (SOTA) methods on our RGB-FIR road dataset. Our methods obtain the two best achievements.

The rest of our paper is organized as follows: the related works are reviewed in Section II. In Section III, we briefly introduce how to get pixel-wisely aligned RGB and FIR images with less ghost phenomena based on the designed hybrid camera. The middle-fusion based model, the architectures of the ECM and HCM models are described in details in Section IV. Section V provides the experiment results on our RGB-FIR road dataset. The quantitative comparisons of different fusion models and the experimental analyses for our ECM and HCM models are also provided in this section. The conclusions are drawn in Section VI. The dataset can be viewed

at https://drive.google.com/drive/folders/1M5UM4XRTV4P_TLsEss0kwaENkHfBoFHi?usp=sharing.

## II. Related Works

### A. Road Detection

In the literature, most road detection approaches depend on RGB cameras. Due to the perspective effect, the parallel road boundaries will intersect to a point (vanishing point) in the image space. Thus, some early road detection methods [14]–[16] concentrate on the detection of vanishing point and the road boundaries. However, there will exist no vanishing point at the T-junction or multiple vanishing points at the fork road cross in the road image. Besides, the road boundaries are not always straight. Though some methods are proposed to fit the road boundary into a complex shape [15], [17], it is hard to obtain an exact road boundary, especially when the obstacles are located on the road region. Some methods assume that the road regions have some similar characteristics based on some hand-crafted features, such as context [16], [18], color [19], [20] and intensity-invariant feature [21]–[23]. Unfortunately, road patches, shadows and water puddles may destroy the characteristic consistency. Moreover, the hand-crafted features cannot extract more delicate information. Due to the development of deep learning and powerful capability in feature extraction, convolutional neural network (CNN) has become a main tool for image-based road detection in recent years, e.g. CN [24], DDN [25], DEEP-DIG [26] and KittiSeg [27]. To further improve performance, some other related tasks are also incorporated into networks. Based on KittiSeg, MultiNet [27] directly segments the road region and detects vehicles simultaneously. RBNet [28] generates the road region and boundaries together. SSLGAN [29] predicts the road region and shape together. Although these methods have achieved good performance on the benchmarks, their performance is always constrained by image quality, and these methods are lack of robustness to variant-illumination conditions.

LiDAR sensor is also a widely used device for road detection. On account of its high accuracy, it is easy to fit a road region based on the output 3D point cloud [30], [31]. But the real road is hard to be treated as a regular 3D shape. Additionally, the original 3D point cloud is not convenient to directly use due to its sparsity and lack of structures. Chen *et al.* [32] reorganize sparse 3D points into a sphere coordinate frame and generate a dense LiDAR imagery. Based on this imagery, Zhang *et al.* [33] propose a scanning based method to detect the road region. Wu *et al.* [34], [35] and Xu *et al.* [36] both build an F-CNN followed by a CRF-like recurrent neural network (RNN) to directly process the LiDAR imagery. Besides, Caltagirone *et al.* [37] just project unstructured 3D points onto the top view and feed the top view image into a fully convolutional network. However, LiDAR points in the top view are very sparse in the depth direction, leading to rougher predictions of road boundaries.

Considering the limitation of each individual sensor, it is preferable to fuse information from both LiDAR and RGB cameras. Earlier approaches [38]–[40] usually first obtain the respective segmentation results from RGB images and LiDAR based on hand-crafted features and traditional classifiers, then fuse the results based on a CRF model. It usually takes these methods much time for feature extraction and CRF computation. Recently, F-CNN has been utilized to obtain fusion results in an end-to-end manner in real time. In PLARD [41], both RGB images and the proposed dense altitude difference images (ADI) are fed into a ResNet-based fusion network. After each ResNet block, the feature maps from ADI images are adaptively fused with the corresponding feature maps from RGB images, and then the fused feature maps are also incorporated into the RGB branch. Fan *et al.* [42] and Wang *et al.* [43] use the RGB image and dense normal vector map as input. The encoder of fusion network is also based on ResNet, while the feature maps from normal vector are also incorporated into the RGB branch after each ResNet block. At last these feature maps at different scales are fed into a dense cascaded decoder for upsampling and prediction. The above two methods both need upsampling for LiDAR data, which restricts the segmentation accuracy. Lv *et al.* [44] also design a two-stream based network where one is for the RGB image in the perspective view and the other is for the artificial features of LiDAR in the bird's eye view, while the perspective feature maps are fused into the bird's eye view for the decoder by a bird's eye transformation [45]. Yang *et al.* [46] create the spatial propagation and transformation to obtain the segmentation in both the perspective and bird's eye views, while the fusion is conducted during the view transformation. However, due to the height difference in the road region, the correspondence between the perspective view and bird's eye one is not quite accurate, which may reduce fusion performance.

### B. Multimodal Semantic Segmentation

Road detection can be treated as a two-class segmentation, thus we also review some recent multimodal semantic segmentation works. FuseNet [47] is a classical RGB-D-based encoder-decoder network, where the encoder includes two branches for each modality, while the feature maps from depth branch are incorporated into the RGB branch after each VGG-16 convolutional block. RFNet [48] also uses two branches in the encoder, where the feature maps from depth branch are incorporated into the RGB branch by using attention-feature complementary modules. In ACNNET [49], the encoder consists of three parallel parts: the RGB, depth and fusion branch. The feature maps in the fusion branch incorporate the outputs of the previous ResNet blocks from the RGB, depth and fusion branches based on the processing of attention complementary module. For polarization based segmentation, Kalra *et al.* [50] use a three-branch encoder. The feature maps after each CNN block from each branch are incorporated by an attention module, and then the output maps in different scales are also fed into a fusion branch based on Mask-RCNN for segmentation and classification. Like [49], EAFNet [51] also use three parallel encoders for RGB, polarization and fusion features, respectively. For RGB-event images, ISSAFE [52] use two branches for each module

in the encoders, while the generated fused feature maps at different scales are fed into one fusion branch as decoder part. In a word, the current multimodal semantic segmentation methods usually use two or more branches in the encoder part, and the feature maps from encoder are incorporated and fed into one fusion branch as decoder for upsampling and segmentation.

### C. FIR Based Road Detection and Semantic Segmentation

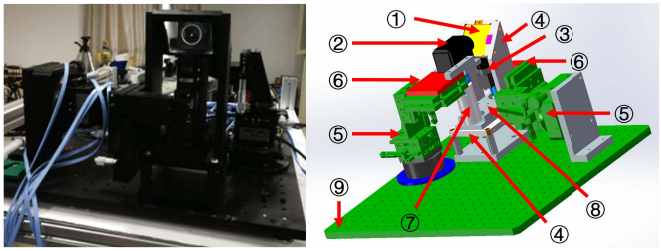
Due to very few benchmark datasets, there are very limited number of works to conduct FIR based road detection or semantic segmentation. Yoon *et al.* propose an online scene-adaptive framework based on FIR image sequences [53]. They suppose that there exists texture consistency for road region in the FIR images, and use a region-growing method to propagate the label to all possible pixels. Next, a coarse-to-fine algorithm propagates the detected region into consecutive frames. However, the consistency of texture and intensity may be affected by road material, shadows, water puddles and lane markings. This makes their results less than those from F-CNN.

The two works most related to ours are MFNet [54] and RTFNet [55], both of which conduct RGB-FIR based semantic segmentation. Based on the lightweight SegNet [56], the MFNet [54] adopts an encoder-decoder structure, which concatenates the feature maps from RGB and FIR branches into a single fusion branch. This work also provides an RGB-FIR dataset, however their labeling does not include the road region. Similarly, RTFNet [55] also uses an encoder-decoder structure, in which the two ResNet-based branches in the encoder are merged into the same fusion branch in the decoder. Nevertheless, the fusion operations are conducted after downsampling blocks in the encoder by adding the FIR feature maps to the RGB feature maps.

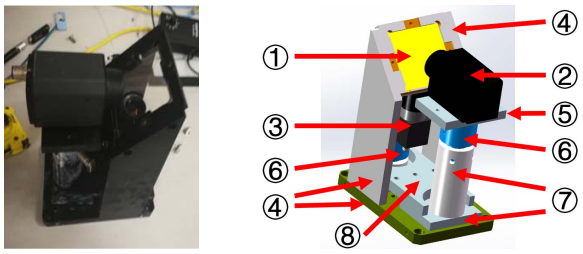
The aforementioned fusion networks incorporate feature maps from different sensors into one fusion branch as decoder for segmentation, while removing the output branches for the respective sensors. This makes the fusion branch neglect some unique useful information extracted only from a single sensor during training. Thus, the fusion network would provide worse prediction results on some validation images than the single sensor based method. Different from these two RGB-FIR approaches and above semantic segmentation methods, our ECM and HCM models recover the individual branches after fusion operation, which are used to imitate the distribution of fusion results during training and improve the segmentation accuracy of the fusion branch.

## III. Image Acquisition

In this section, we first introduce our RGB-FIR hybrid image device briefly. For more details, it is suggested to refer to our previous work [2]. Our hybrid imaging device consists of a visible-light camera and a thermal one, illustrated in Figure 2. We also exploit a beam splitter to keep the two optical-axes coaxial and make the optical centers approximately meet at the same point. Thereafter, there is a pixel-wise correspondence between RGB and FIR images, which is




(a) Adjuster platform: ① beam splitter, ② FIR camera, ③ RGB camera, ④ frame, ⑤ adjuster, ⑥ camera base, ⑦ support, ⑧ fastener, ⑨ optical table




(b) RGB-FIR hybrid camera: ① beam splitter, ② FIR camera, ③ RGB camera, ④ frame, ⑤ camera base, ⑥ universal joint, ⑦ support, ⑧ fastener

Fig. 2. Adjuster platform for hybrid camera and RGB-FIR hybrid camera.

subject to a homography transformation based on the stereo calibration, achieving the pixel-wise alignment of RGB and FIR frames.

Even though a homography matrix can provide very accurate correspondence in theory, one problem still remains for image acquisition, i.e. the type of shutter adopted for exposure. In practice, a large amount of RGB cameras are based on the global shutter, which captures the whole image at the same time. Whereas, most of the existing FIR cameras are based on the rolling shutter, in which the top of the image is obtained earlier than the bottom. It means that, if the RGB and FIR cameras start the exposure at the same time, the bottom part of RGB image cannot perfectly overlay the bottom part of FIR image, especially when either the hybrid camera or objects are moving in the scene. This can arouse the "ghost" phenomenon, just like a ghost haunting near a person.

Because we do not know the exact exposure details on how the rolling shutter works in a consumer-grade FIR camera (like the one we used), it is almost impossible for us to fully eliminate the ghost phenomena. Instead, we expect to attenuate the ghost phenomena in the road region as much as possible. For this purpose, we postpone triggering the RGB camera until the FIR camera captures the road region in the bottom. This can significantly alleviate the ghost phenomena on the road region.

Figure 3(a) gives an example of the MFNet dataset [54], which is captured by an RGB-FIR camera, InfReC R500, which does not adopt a coaxial style and synchronization in image capturing. For a distant static object (e.g. buildings in the image), since its distance to the RGB-FIR camera is much larger than the baseline between RGB and FIR cameras, the disparities between corresponding pixels are close to zero. Nevertheless, there exists large misalignment when an object is close to the moving camera, just as the car in the left and the billboard in the right.

(a) An example from the MFNet dataset [54].

(b) An example from the KAIST pedestrian dataset [3].

(c) An example from our RGB-FIR road detection dataset.

Fig. 3. The example images from the current three RGB-FIR datasets. Left: RGB images, Middle: FIR images, Right: blending images in which the red channel corresponds to the grayscale of RGB image and the green channel is the FIR image.

Figure 3(b) is selected from the KAIST pedestrian dataset [3], whose cameras are aligned in a coaxial way although the ghost phenomena are not carefully taken care of. By virtue of the specific pose of their RGB and FIR cameras, the ghost phenomena in the left part of images are worse than those in the right part. Thus, the ghost phenomena obviously appear at the pedestrians in the left boundary, and those in the FIR image cannot match the ones in the RGB image. This also makes the dataset not so suitable for road detection.

Figure 3(c) shows an example acquired by our hybrid camera with the strategy of ghost phenomenon attenuation. The image was taken when the vehicle was turning quickly at a T-junction. Although the ghost phenomena appear on the traffic lights in the top of the image, thanks to the trigger delay of the RGB camera, the road markers in the RGB image still well match those in the FIR image. This indicates that our strategy could effectively attenuate ghost phenomena in road region.

## IV. The Proposed Network

Based on our RGB-FIR hybrid camera, we can capture pixel-wisely aligned RGB and thermal images simultaneously. Intuitively, the information from RGB and FIR images supplements each other. RGB images provide richer color and clearer texture information in good illumination conditions, while FIR images contain more information in adverse illumination conditions. Based on the lightweight F-CNN ERFNet [57], the middle-fusion based model, our ECM and HCM models are proposed to gradually promote the fusion of features and achieve better performance.

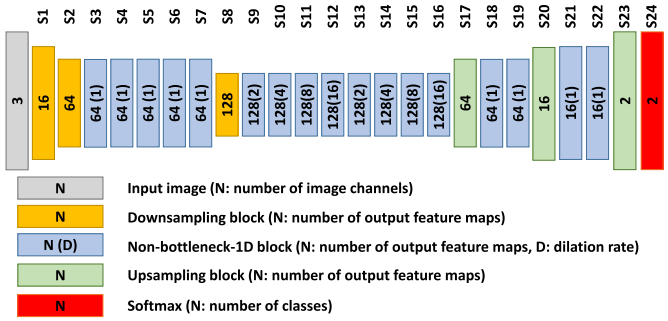



Fig. 4. The architecture of ERFNet.

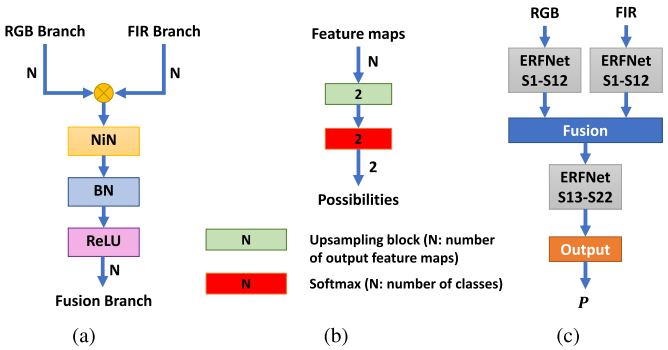



Fig. 5. (a) Fusion block, (b) Output block, (c) middle-fusion based ERFNet (ERFNet+MF).

### A. The Basic Model and Middle Fusion

The ERFNet [57] is a lightweight architecture and provides accurate semantic segmentation results in real time, which is suitable for autonomous driving. The architecture is shown in Figure 4. For illustrative purposes, each stage is assigned to an index. The whole network consists of three parts: downsampling block, non-bottleneck-1D block and upsampling block. Among these blocks, the non-bottleneck-1D block is the key module of ERFNet. Unlike the general residual module, each $3 \times 3$ convolution kernel is replaced with two 1D kernels: one is $3 \times 1$, and the other is $1 \times 3$. To obtain more contexts, the latter pair of 1D kernels are changed to the dilated ones. The name "non-bottleneck" means that all the convolutions do not change the number of feature channels. At last, all these blocks are stacked sequentially to construct an encoder-decoder structure, which generates the segmentation results of the same size as the input images.

Though RGB or FIR images could be processed by the ERFNet individually, fusing RGB and FIR features may intuitively provide complementary information for road detection with different illumination conditions. Here, we just choose the middle fusion (MF) as the concatenation based fusion model, where the fusion operation is conducted after the $S12$ stage of the ERFNet model (i.e. the middle of the basic model). To maintain the symmetry of each branch, and to keep the same number of parameters in different branches, the single-channel FIR image is converted to a 3-channel image. The details of the fusion model are displayed in Figure 5. To keep the same dimensions of input feature maps as those of output, and to preserve the structure of ERFNet and the quantity of parameters, the fusion block first concatenates

the feature maps together, doubling the number of feature maps. Then a Network-in-Network (NiN) is introduced after the concatenation to reduce the number of feature maps to the original input, followed by the batch normalization (BN) and the ReLU activation. Besides, for ease of illustration, the stages of $S23$ and $S24$ in the ERFNet are treated as the output block (Figure 5(b)).

Let regard a combination of $3 \times 1$ and $1 \times 3$ convolution as a "complete" $3 \times 3$ convolution. Actually, the original image is subject to at least 18 layers of complete $3 \times 3$ convolution before the fusion block, which exceeds those in the VGG19 network [58]. Thus, we could argue that the generated feature maps after the $S12$ stage involve semantic meaning. Meanwhile, as the middle-level features, they still retain some fine details for semantic segmentation. In addition, for the middle fusion, the fused feature maps are still subject to about 16 layers of complete $3 \times 3$ convolution, which is able to mix the RGB and FIR feature fully. Referring to the RGB-FIR based pedestrian detection approaches [4], [5], [59], their experimental results also demonstrate that the middle-level fusion can improve detection performance. Consequently, we just choose the middle-fusion based ERFNet model as the concatenation based fusion model.

### B. The Extended Cross Model

In general, the target of semantic segmentation network is to minimize the cross-entropy loss function, which measures the probability differences between the estimated results and the ground-truth. The cross-entropy loss is written as

$$L_{seg}(y, P) = -\frac{1}{H \times W} \cdot \sum_{\mathbf{x}} \sum_{i=1}^{C} y_{(\mathbf{x},i)} \cdot \log P_{(\mathbf{x},i)}, \quad (1)$$

where $y_{(\mathbf{x},i)}$ is the ground-truth label for the $i$-th class at the pixel $\mathbf{x}$, and $P_{(\mathbf{x},i)}$ is the estimated probability for the $i$-th class at the pixel $\mathbf{x}$. $C$ is the number of class. $H$ and $W$ correspond to the height and width of the output image, respectively. For road detection, the ground-truth consists of two classes: road region and background, thus $C = 2$. Nevertheless, due to the preserved unique information in the individual branch, this loss function does not consider the difference and complementarity between the RGB and FIR branches. As a result, The degree of fusion is not sufficient.

In [13], Jaritz *et al.* propose a new architecture to perform cross-model learning and domain adaptation for 3D semantic segmentation. Different from the other fusion networks using diverse types of fusion layers, they just use two detached branches. Besides, the output feature maps after 2D-3D projection in each branch are sent to two distinct heads, an original output head and a mimic one. Each head includes a linear layer with a ReLU followed by a softmax function, which generates respective probabilities. For these two probabilities, one represents the distribution of segmentation results from the current branch, while the other is utilized to imitate the distribution from the other branch. Then an auxiliary loss is defined as the KL divergence to regularize the relationship between the original probability distribution from the current branch and the mimic one from the other branch as follows,

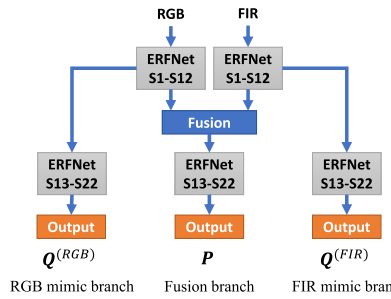



Fig. 6. The extended cross model for middle fusion (ERFNet+MF+ECM).

$$L_{xM}(P, Q) = D_{KL}(P||Q) = -\frac{1}{H \times W} \cdot \sum_{\mathbf{x}} \sum_{i=1}^{C} P_{(\mathbf{x},i)} \cdot \log \frac{P_{(\mathbf{x},i)}}{Q_{(\mathbf{x},i)}}, \quad (2)$$

where $P_{(\mathbf{x},i)}$ is the output probability from the current branch, and $Q_{(\mathbf{x},i)}$ is the mimic probability from the other branch. Moreover, they also give a fusion version, in which the two original heads for each branch are replaced by two layers. The first layer concatenates the projected feature maps from two branches and the second one is a linear layer receiving the feature maps as input with a ReLU followed by a softmax function. Meanwhile, the mimic heads in each branch are still reserved, and the $P_{(\mathbf{x},i)}$ in (2) is replaced with the output of fusion head.

The cross-model based fusion not only obtains the shared information between two branches (i.e. fusion head), but also preserves the private information for each branch (i.e. mimic heads), thus achieving superior performance for 3D semantic segmentation. However, two drawbacks still exist and restrict the model's application. First, in 3D semantic segmentation models, the fusion usually operates at the late stage after 2D-3D projection (i.e. late fusion), where the feature maps from each branch have already contained much discrimination information. But for image based segmentation method, fusion usually operates at an earlier stage, such as the middle fusion, where the feature maps from each branch have less discrimination information than those from late fusion. It is unclear whether the mimic head provides sufficient discrimination information. In addition, because of the existence of downsampling and upsampling operations, the output size of feature maps changes at different stages in the image-based semantic segmentation model. It is inappropriate to use a simple mimic head (i.e. a linear layer with ReLU followed by softmax) to preserve private information.

To solve these problems, we propose a new architecture to extend the original cross model to the image-based semantic segmentation, called the extended cross model (ECM). Figure 6 illustrates its structure. It depends on the middle-fusion model. The size of feature map input into the fusion block is only 1/8 of the original image size. To guarantee that the output size of the mimic branch is equal to that of the fusion branch and to maintain the consistency across the fusion branch and two mimic ones, both two mimic branches and the fusion branch utilize the same structure. For the layers

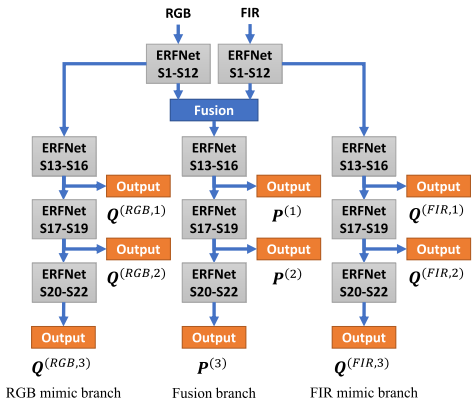



Fig. 7. The hierarchical cross model for middle fusion (ERFNet+MF+HCM).

from $S17$ to $S23$, feature maps are gradually enlarged to the original size of the input image, while the feature maps from the later stages gradually gain more discrimination capability in theory. This enables the mimic branch to well imitate the probability distribution of the fusion branch, and achieves better performance for the fusion network after training. Hence, the loss function of extended cross model is formulated as

$$\begin{aligned} &L_{ECM}(P, Q^{(RGB)}, Q^{(FIR)}) \\ &= D_{KL}(P||Q^{(RGB)}) + D_{KL}(P||Q^{(FIR)}) \\ &= -\frac{1}{H \times W} \cdot \sum_{\mathbf{x}} \sum_{i=1}^{C} P_{(\mathbf{x},i)} \cdot \left( \log \frac{P_{(\mathbf{x},i)}}{Q^{(RGB)}_{(\mathbf{x},i)}} + \log \frac{P_{(\mathbf{x},i)}}{Q^{(FIR)}_{(\mathbf{x},i)}} \right), \end{aligned} \tag{3}$$

where $P_{(\mathbf{x},i)}$ is the output probability of the fusion branch for the $i$-th class. The $Q^{(RGB)}_{(\mathbf{x},i)}$ is the mimic probability of the RGB mimic branch for the $i$-th class. The $Q^{(FIR)}_{(\mathbf{x},i)}$ is the mimic probability of the FIR mimic branch for the $i$-th class. Then the whole loss function of our ECM is

$$\begin{aligned} Loss_{ECM} &= L_{seg}(y, P) + \lambda_c \cdot L_{ECM}(P, Q^{(RGB)}, Q^{(FIR)}) \\ &= -\frac{1}{H \times W} \cdot \sum_{\mathbf{x}} \sum_{i=1}^{C} \Bigg[ y_{(\mathbf{x},i)} \cdot \log P_{(\mathbf{x},i)} \\ &\quad + \lambda_c \cdot P_{(\mathbf{x},i)} \cdot \left( \log \frac{P_{(\mathbf{x},i)}}{Q^{(RGB)}_{(\mathbf{x},i)}} + \log \frac{P_{(\mathbf{x},i)}}{Q^{(FIR)}_{(\mathbf{x},i)}} \right) \Bigg], \end{aligned} \tag{4}$$

where $\lambda_c$ is the hyperparameter weight of $L_{ECM}$.

### C. The Hierarchical Cross Model

During the encoding process, the upsampled feature maps often lack detail information because it can be discarded by the downsampling operations in the encoder part. Consequently, an auxiliary loss is usually added to the stage where upsampling occurs, which constitutes a hierarchical loss (HL) function to ensure the recovery of details. That means we will obtain several segmentation results at diverse resolutions in the fusion branch. Similarly, the extended cross model in Figure 6 could be also transformed to a hierarchical version. Both the fusion and mimic branches have several upsampling operations, and to reserve the details from the mimic branches, each mimic branch also constitutes its own hierarchical loss. This model is called the "hierarchical cross model" (HCM), shown in Figure 7. Here all the mimic branches own a series of auxiliary outputs with different scales to imitate the probability distribution from the fusion branch at different stages, Thus, the whole loss function of our HCM is

$$\begin{aligned} Loss_{HCM} &= \sum_{l=1}^{L} \Big\{ L_{seg}(y^{(l)}, P^{(l)}) \\ &\quad + \lambda_c \cdot L_{ECM}(P^{(l)}, Q^{(RGB,l)}, Q^{(FIR,l)}) \Big\} \\ &= -\sum_{l=1}^{L} \Bigg\{ \frac{1}{H^{(l)} \times W^{(l)}} \cdot \sum_{\mathbf{x}} \sum_{i=1}^{C} \Bigg[ y^{(l)}_{(\mathbf{x},i)} \cdot \log P^{(l)}_{(\mathbf{x},i)} \\ &\quad + \lambda_c \cdot P^{(l)}_{(\mathbf{x},i)} \cdot \left( \log \frac{P^{(l)}_{(\mathbf{x},i)}}{Q^{(RGB,l)}_{(\mathbf{x},i)}} + \log \frac{P^{(l)}_{(\mathbf{x},i)}}{Q^{(FIR,l)}_{(\mathbf{x},i)}} \right) \Bigg] \Bigg\}, \end{aligned} \tag{5}$$

where $L$ is the number of hierarchies for upsampling. The $H^{(l)}$ and $W^{(l)}$ are the height and width of the output image at the $l$-th upsampling block, respectively. The $y^{(l)}_{(\mathbf{x},i)}$ denotes the ground-truth label for the $i$-th class at the pixel $\mathbf{x}$ on the $l$-th output hierarchy. The $P^{(l)}_{(\mathbf{x},i)}$ represents the output probability of the fusion branch for the $i$-th class on the $l$-th output hierarchy. The $Q^{(RGB,l)}_{(\mathbf{x},i)}$ is the mimic probability of the RGB mimic branch for the $i$-th class on the $l$-th output hierarchy, while $Q^{(FIR,l)}_{(\mathbf{x},i)}$ is the mimic probability of the FIR mimic branch for the $i$-th class on the $l$-th output hierarchy. Specifically, we set $L = 3$ in our ERFNet based HCM model.

## V. Experiments on Our RGB-FIR Road Dataset

### A. RGB-FIR Road Dataset

Based on the built RGB-FIR camera, we have acquired 500k pixel-aligned RGB-FIR image pairs at the rate of 29.97Hz (the fixed frequency of FIR camera). These images are captured in Nanjing, China, and include two major categories of scenes, urban and suburban/village regions. To evaluate our models, we select 2036 image pairs as the RGB-FIR road detection dataset, and manually annotate ground-truth. In this dataset, 1035 image pairs are captured in the daytime including morning, afternoon and dusk, while the other 1001 image pairs are captured at night. Some examples are shown in Figure 1.

When annotating the images, we divide them into three parts, the images captured in good illumination conditions, the ones with partially blurred regions (e.g., shadows) and the ones captured in adverse illumination conditions. Because the RGB images contain much more texture information and details than the FIR images, for the first parts of images, we directly annotate the road region on the RGB images. Since our hybrid camera could provide pixel-wisely aligned RGB-FIR images, these annotations are also the ground-truth of the corresponding FIR images. Next, we just project the labeled road regions onto the FIR images, and analyze how the road regions and the road boundaries are presented in the FIR images. With this prior knowledge, we then label the remaining two parts. For the images with partially blurred

TABLE I
THREE SPLITS FOR THE RGB-FIR ROAD DETECTION DATASET

| Split | Training | Validation | Testing |
|---|---|---|---|
| Day | 588 | 223 | 224 |
| Night | 597 | 199 | 205 |

regions, we first annotate on the RGB images with image adjustment operations (for example, brightness control and sharpening) to determine the road region as much as possible. Then we fine-tune the road boundaries in the FIR images. For the images captured in adverse illumination conditions, we first annotate on the FIR images since the FIR images could show the whole road region. Because there are also FIR images with low contrast, the image adjustment operations on FIR images are also needed. After labeling on the FIR image, we project the road region onto the corresponding RGB image and fine-tune the boundary, especially near the vehicles, sticks and guard barriers. To obtain more details, the RGB images are also needed to conduct image adjustment. Based on the aforementioned strategies, we could guarantee the consistency of ground-truth between RGB and FIR images.

For evaluation purposes, 1185 image pairs are regarded as the training set, and 422 image pairs are divided into the validation set, and the rest 429 image pairs belong to the testing set. More details about the dataset are provided in Table I. The resolution of both RGB and FIR images is $640 \times 480$.

### B. Experimental Settings

During the training process, all our designed networks for evaluation are conducted based on the same operations and settings as follows. First, these networks are implemented in the PyTorch framework and the whole experiments are conducted on a server with a single GPU of NVIDIA GeForce RTX 2080 Ti. Second, both the encoder and decoder are trained from scratch simultaneously. During training, the batch size is set to 4 and the number of epochs is 300. The Adam [60] optimization method is adopted during training with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate decreases depending on the poly learning policy [61] as

$$\eta_i = \eta_0(1 - \frac{i}{N})^{\alpha}, \tag{6}$$

where $i$ is the number of current epoch, and $N$ is the total number of epochs and set to 300. The initial learning rate $\eta_0 = 0.0005$ and $\alpha = 0.9$. To avoid overfitting, $l2$ regularization is also applied and the weight decay parameter is set to 0.0001. As the regularization hyper-parameter in the loss functions (4) and (5), $\lambda_c$ is set to 0.1. Besides, for the KL divergence in the loss function, we detach the output of fusion branch from the mimic branches. For data augmentation, we just utilize horizontal flipping and random translation with the horizontal and vertical offset no more than 2 pixels, as the same settings in ERFNet [57].

To evaluate the performance of the segmentation results, four measurement metrics are provided [62], [63], i.e. precision (PRE), recall (REC), F-score and intersection over union (IoU). The formulas of these four metrics are listed as following:

$$PRE = \frac{TP}{TP + FP} \tag{7}$$

$$REC = \frac{TP}{TP + FN} \tag{8}$$

$$F - score = \frac{2 * PRE * REC}{PRE + REC} \tag{9}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

where TP means true positive, FP representing false positive and FN denoting false negative. Thus, the weights with largest IoU score on the validation set are chosen for evaluation on the testing data.

### C. Experimental Results

We evaluate the fusion methods, i.e. the middle-fusion based ERFNet model (ERFNet+MF), the ERFNet+MF model with extended cross model (ERFNet+MF+ECM), and the ERFNet+MF model with hierarchical cross model (ERFNet+MF+HCM). We compare these methods with two base methods. The first base method only uses RGB image as the input of the original ERFNet [57], and is denoted by the "ERFNet+RGB". The second base method uses 3-channel FIR image as the input of the ERFNet, denoted the "ERFNet+FIR", where each channel of the FIR image is the same. We also compare with an ERFNet-like method with a 4-channel input composed of a 3-channel RGB and an 1-channel FIR image, called the "ERFNet-like+4-channel". Moreover, our methods are compared with two SOTA RGB-FIR fusion methods MFNet [54] and RTFNet50 [55], one typical image-based road detection method KittiSeg [27], one SOTA RGB-normal fusion based road detection method SNE-RoadSeg-50 [42], and a recent efficient RGB-D semantic segmentation method RFNet [48]. Specially, the 3-channel normal vector map in the SNE-RoadSeg-50 model is replaced by the 3-channel FIR image.

Table II(a) shows the overall performance on the whole testing set. Our middle-fusion based ERFNet model and its extensions with the ECM and HCM all obtain better evaluation results than the two base methods. This indicates that the fusion of RGB and FIR images could enhance network performance. The IoU value of the ERFNet+MF is 96.88%, while those from the ERFNet+MF+ECM and ERFNet+MF+HCM increase by 0.25% and 0.37%, respectively. This proves that our ECM and HCM models are effective. Besides, the ERFNet+MF+ECM and ERFNet+MF+HCM also achieve better performance compared with the MFNet and RTFNet50, also demonstrating the superiority of our models. The 4-channel input method ERFNet-like+4-channel obtains the IoU value of 96.43%, less than the middle-fusion model. This indicates that the method with a direct 4-channel input can not adequately utilize complementary information between RGB and FIR images. Compared with the VGG-16 based KittiSeg, the ERFNet with a deeper structure could achieve better performance, no matter what the input is. Whereas, the

TABLE II

EVALUATION ON THE TESTING SET OF THE RGB-FIR ROAD DETECTION DATASET FOR DIFFERENT METHODS

(a) All Data

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+RGB [57] | 96.35% | 97.30% | 96.82% | 93.84% |
| ERFNet+FIR [57] | 97.06% | 98.21% | 97.63% | 95.37% |
| MFNet [54] | 97.68% | 98.06% | 97.87% | 95.83% |
| RTFNet50 [55] | **98.39%** | 98.14% | 98.26% | 96.58% |
| KittiSeg+RGB [27] | 95.56% | 96.75% | 96.15% | 92.59% |
| KittiSeg+FIR [27] | 96.04% | 98.26% | 97.14% | 94.43% |
| SNE-RoadSeg-50 [42] | 97.15% | 97.54% | 97.35% | 94.83% |
| ERFNet-like+4-channel | 98.04% | 98.33% | 98.18% | 96.43% |
| RFNet [48] | 97.97% | 98.53% | 98.25% | 96.56% |
| ERFNet+MF | 97.88% | **98.96%** | 98.42% | 96.88% |
| ERFNet+MF+ECM | 98.19% | 98.90% | 98.54% | 97.13% |
| ERFNet+MF+HCM | 98.28% | 98.93% | **98.61%** | **97.25%** |

(b) Daytime Subset

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+RGB [57] | 97.67% | 98.17% | 97.92% | 95.92% |
| ERFNet+FIR [57] | 96.34% | 97.07% | 96.70% | 93.62% |
| MFNet [54] | 97.34% | 97.41% | 97.38% | 94.89% |
| RTFNet50 [55] | 98.38% | 97.14% | 97.76% | 95.61% |
| KittiSeg+RGB [27] | 97.44% | 97.87% | 97.65% | 95.41% |
| KittiSeg+FIR [27] | 94.45% | 97.29% | 95.85% | 92.03% |
| SNE-RoadSeg-50 [42] | 96.51% | 96.38% | 96.44% | 93.13% |
| ERFNet-like+4-channel | 97.97% | 97.55% | 97.76% | 95.62% |
| RFNet [48] | 97.91% | 97.81% | 97.86% | 95.81% |
| ERFNet+MF | 97.78% | **98.52%** | 98.15% | 96.37% |
| ERFNet+MF+ECM | 98.22% | 98.51% | 98.36% | 96.78% |
| ERFNet+MF+HCM | **98.53%** | 98.47% | **98.50%** | **97.04%** |

(c) Nighttime Subset

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+RGB [57] | 95.10% | 96.48% | 95.79% | 91.91% |
| ERFNet+FIR [57] | 97.74% | 99.30% | 98.51% | 97.07% |
| MFNet [54] | 98.00% | 98.68% | 98.34% | 96.73% |
| RTFNet50 [55] | 98.39% | 99.09% | 98.74% | 97.51% |
| KittiSeg+RGB [27] | 93.80% | 95.68% | 94.73% | 89.99% |
| KittiSeg+FIR [27] | 97.58% | 99.18% | 98.37% | 96.80% |
| SNE-RoadSeg-50 [42] | 97.76% | 98.65% | 98.20% | 96.47% |
| ERFNet-like+4-channel | 98.10% | 99.08% | 98.59% | 97.22% |
| RFNet [48] | 98.02% | 99.22% | 98.62% | 97.27% |
| ERFNet+MF | 97.97% | **99.38%** | 98.67% | 97.38% |
| ERFNet+MF+ECM | **98.16%** | 99.28% | **98.72%** | **97.47%** |
| ERFNet+MF+HCM | 98.05% | **99.38%** | 98.71% | 97.46% |

SNE-RoadSeg-50 does not obtain a desirable result. We guess this is because the processing and fusion strategy for normal vector is not suitable for FIR images. The F-measure and IoU scores of RFNet is much larger than ERFNet-RGB and ERFNet-FIR, which validates the effectiveness of RGB and FIR fusion. However, its performance is still weaker than our ERFNet-based fusion model. One reason is that its backbone is shallower, which is hard to extract more semantic information for the RFNet. Another reason is its simple upsampling block. Its upsampling block only consists of an interpolation operation and a convolution layer with pre-operations of batch normalization and ReLU activation. This might not be enough to recover the details of the upsampled result, even with the help of skip connections from the encoder. In addition, all the metrics of the ERFNet+FIR are larger than those of the ERFNet+RGB. Since FIR images are less illumination-dependent than RGB images during night, these results directly reflect the decreasing segmentation capability for RGB-image based method at nighttime.

Table II(b) and (c) report the performance on the daytime and nighttime subset in the testing set, respectively. First, compared with the two base methods, the values of F-score and IoU of the ERFNet+MF, ERFNet+MF+ECM and ERFNet+MF+HCM are larger on both the daytime and nighttime data. This proves the effectiveness of our models in fusing RGB and FIR information. Second, all the metrics of the ERFNet+RGB on the daytime data are much better than those of the ERFNet+FIR. It indicates that abundant color and texture information is crucially important for road detection. In contrast, the ERFNet+FIR achieves preferable performance on the nighttime data. This implies that the relatively simple textures from FIR images could still offer valuable semantic information for road region. Although dark environment makes the RGB camera capture insufficient discrimination information, the ERFNet+RGB method still acquires the IoU of 91.91% and the F-score of 95.79%, even on the nighttime data. This shows that the captured RGB images still contain useful features of the road region, which are perceivable to the algorithm although invisible to human vision. Third, all these RGB-FIR fusion methods acquire higher scores of IoU and F-score on the nighttime data than those on the daytime data. This can be explainable in two perspectives. (1) The traffic condition in the daytime is relatively complex. Traffic jams are easy to occur during the day, while the road has less traffic at night. (2) The illumination condition in the daytime is also more complex than that at night. For example, sunlight can cast dappled shadows on road in both RGB and FIR images during daytime, while car light or street lamp cannot in FIR images. In addition, the ERFNet+MF+HCM obtains a slightly lower IoU score than the ERFNet+MF+ECM on the nighttime subset. This is because the trained model with the best IoU score on the whole validation set is chosen as the model to be tested. Thus, even though the HCM model obtains higher F-measure and IoU scores on the whole testing set, it is possible that some metrics of the HCM model is slightly lower than those of the ECM model on a subset of testing images.

Figure 8 shows the qualitative results of all the compared methods in Table II. Compared with the ERFNet+RGB and ERFNet+FIR, the ERFNet+MF can utilize the information from both RGB and FIR images to overcome the effect from variant illumination environment, whilst the ERFNet+MF+ECM and ERFNet+MF+HCM could further achieve superior performance. For example, in Figure 8, although the plaques, water or shadows on the road in the 1st, 2nd, 7th and 8th row affect the road imaging in both the RGB and FIR images, which would disturb the single-sensor based method, the fusion based methods still provide good segmentation results. In the 3rd, 4th, 9th, 10th and 11th row, the vehicles stopping for a while may make their underneath road region a little warmer, and the corresponding regions in FIR images are brighter. This may reduce the accuracy

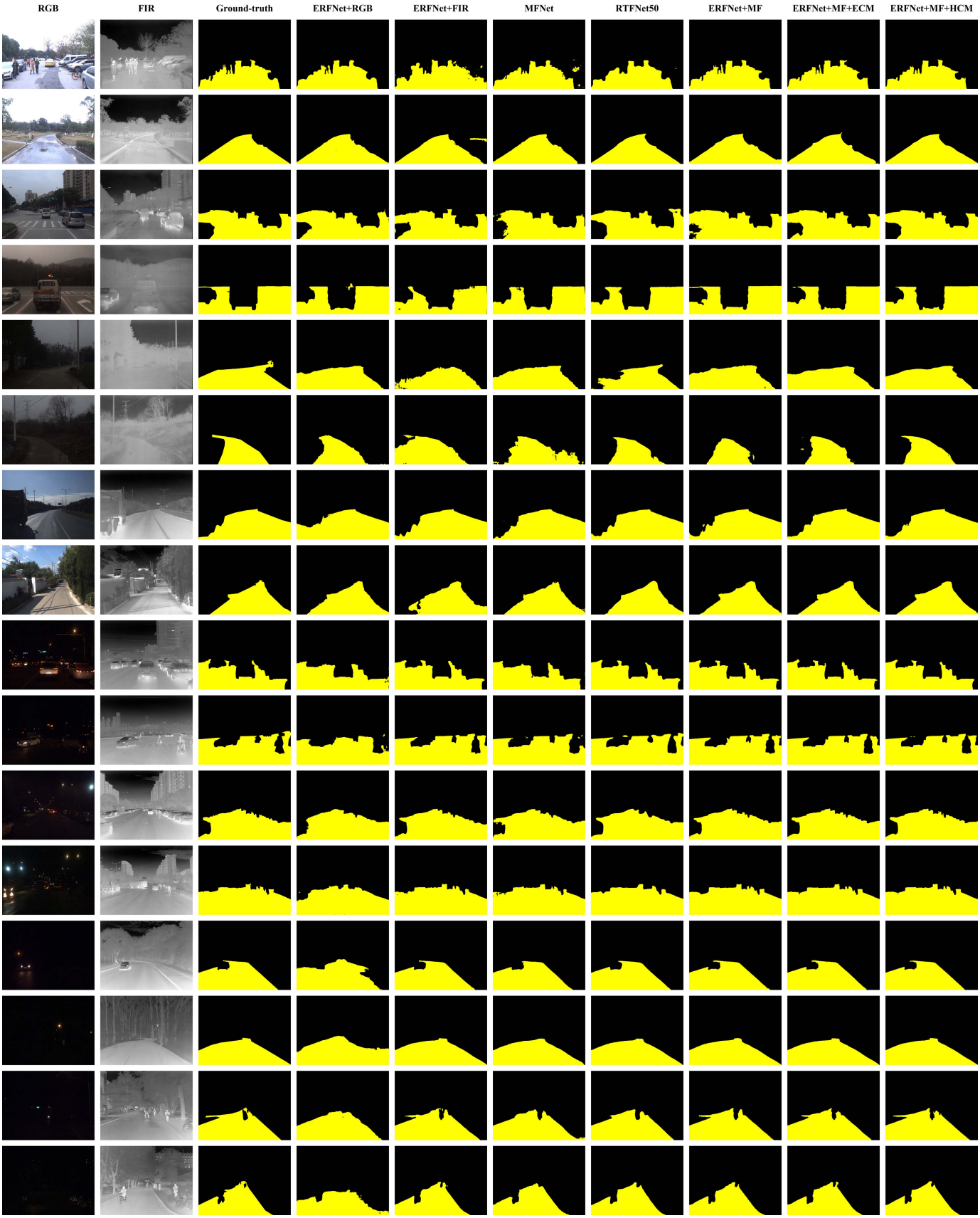



Fig. 8. Qualitative comparisons on RGB-FIR road dataset.

of the single-FIR based method. However, our ECM and HCM models still recognize these regions and predict desired boundaries around the vehicles. In the 12th and 13th row, the lights from cars are reflected off the ground and the generated

light spots affect the RGB images. Nevertheless, the FIR and fusion-based methods could work well in this condition. For these fuzzy and dark RGB images in the 5th, 6th, 13th, 14th, 15th and 16th row, the ECM and HCM models also predict the road regions with fine boundaries.

The fusion branch could not only learn fused features but also more unique information from the mimic branches. According to (4) and (5), it still needs to keep a trade-off among the fusion, RGB and FIR branches during training. When RGB and FIR image provide contrary information, the fusion branch may be influenced and predict a wrong segmentation result. Figure 9 shows some typical negative results. In the 1st row, the soil on the right bottom is wrongly recognized as a road region by the FIR-based method. Thus, all the fusion methods including RTFNet50 also misclassify this region. In the 5th row, the grassland on the right bottom is wrongly recognized as a road region by the RGB-based method, whilst all the fusion methods also provide wrong predictions. In the 7th row, the left bottom region is wrongly classified as a road region by the FIR-based method, and this region is not recognized correctly more or less by these fusion methods. In the 8th row, it is hard to draw a distinction between the wall and road region in the FIR image, which makes the FIR-based methods obtain poor results. Moreover, when RGB and FIR images provide misleading information at the same time, the fusion branch is hard to correct it. The shadow near the barrier in the 3rd row makes all the methods treat the shadow region as a non-road region. On the contrary, the sidewalk and step under the gallery in the 4th row are regarded as a road region by all these methods. Besides, the small and hazy objects in both RGB and FIR images may be neglected, e.g., the traffic cone on the right of the image in the 2nd row, and the small road region between two pedestrians in the middle of the image in the 6th row.

### *D. Execution Profile*

We also compare the frame rate, the number of parameters and FLOPs for our models and two RGB-FIR based methods in Table III. All these methods are measured with a GPU of NVIDIA GeForce RTX 2080 Ti. Our ERFNet+MF+ECM and ERFNet+MF+HCM models have 5.139 million and 5.144 million parameters, respectively. Nevertheless, the mimic branches and the auxiliary outputs for hierarchical loss are not applicable during the testing process. By simplifying these parts, the number of parameters is reduced to the one of the ERFNet+MF, while the frame rate is also close to that of the ERFNet+MF. Compared with the very lightweight network MFNet, our methods achieve superior performance. Compared with the RTFNet50, our proposed methods are faster, while having fewer parameters and less computation.

### *E. Ablation Studies*

*1) Components in HCM Model:* As the HCM consists of several components, to evaluate their effectiveness, several methods with different components are compared with the ERFNet+MF+HCM (Figure 7). These methods include the middle-fusion based ERFNet (ERFNet+MF, Figure 5(c)), the ERFNet+MF with a hierarchical loss on the fusion branch (ERFNet+MF+HL, Figure 10(a)), the ERFNet+MF based ECM (ERFNet+MF+ECM, Figure 6) and the ERFNet+MF+ECM with a hierarchical loss on the fusion branch (ERFNet+MF+ECM+HL, Figure 10(b)). Table IV depicts the evaluation results. As can be seen, compared with the basic ERFNet+MF model, adding these components promotes the network performance gradually. With the addition of a hierarchical loss on the fusion branch, the IoU of the ERFNet+MF+HL increases by 0.13%. Similarly, the IoU of the ERFNet+MF+ECM+HL is also 0.06% larger than that of the ERFNet+MF+ECM. At last, the ERFNet+MF+HCM achieves a 0.37% larger IoU value than the basic model.

*2) Mimic Branch:* Considering the complexity of decoder, the structure of mimic branch also impacts the detection results. Here we also compare the ERFNet+MF+ECM with its variant, the ERFNet+MF+ECM(8x). Its architecture is shown in Figure 11, where three 2x upsampling blocks and several non-bottleneck blocks followed by the softmax function in the mimic branches are simplified to a simple 8x upsampling block followed by the softmax function. The block "Output (8x)" in Figure 11 means that the original 2x upsampling block in the output block (Figure 5(b)) is directly replaced by an analogous 8x upsampling block. The results are notified in Table V. The simple 8x mimic branch still preserves the effectiveness, but its IoU value is less than

TABLE III
COMPARISONS ON FRAME RATE, PARAMETERS AND FLOPS

| Method | Frame rate (fps) | Parameters (million) | FLOPs (billion) |
|---|---|---|---|
| ERFNet [57] | 63.36 | 2.063 | 17.271 |
| MFNet [54] | 95.90 | 0.734 | 7.764 |
| RTFNet50 [55] | 36.08 | 185.233 | 244.165 |
| ERFNet+MF | 41.68 | 3.180 | 26.337 |
| ERFNet+MF+ECM | 26.41 | 5.139 | 43.065 |
| ERFNet+MF+ECM (w/o mimic branches & auxiliary outputs) | 40.84 | 3.180 | 26.337 |
| ERFNet+MF+HCM | 25.68 | 5.144 | 43.243 |
| ERFNet+MF+HCM (w/o mimic branches & auxiliary outputs) | 40.97 | 3.180 | 26.337 |

TABLE IV
COMPARISONS ON FUSION BRANCHES

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+MF | 97.88% | **98.96%** | 98.42% | 96.88% |
| ERFNet+MF+HL | **98.35%** | 98.62% | 98.48% | 97.01% |
| ERFNet+MF+ECM | 98.19% | 98.90% | 98.54% | 97.13% |
| ERFNet+MF+ECM+HL | 98.22% | 98.94% | 98.58% | 97.19% |
| ERFNet+MF+HCM | 98.28% | 98.93% | **98.61%** | **97.25%** |

TABLE V
COMPARISONS ON MIMIC BRANCHES

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+MF | 97.88% | **98.96%** | 98.42% | 96.88% |
| ERFNet+MF+ECM(8x) | 98.13% | 98.74% | 98.43% | 96.92% |
| ERFNet+MF+ECM | **98.19%** | 98.90% | **98.54%** | **97.13%** |

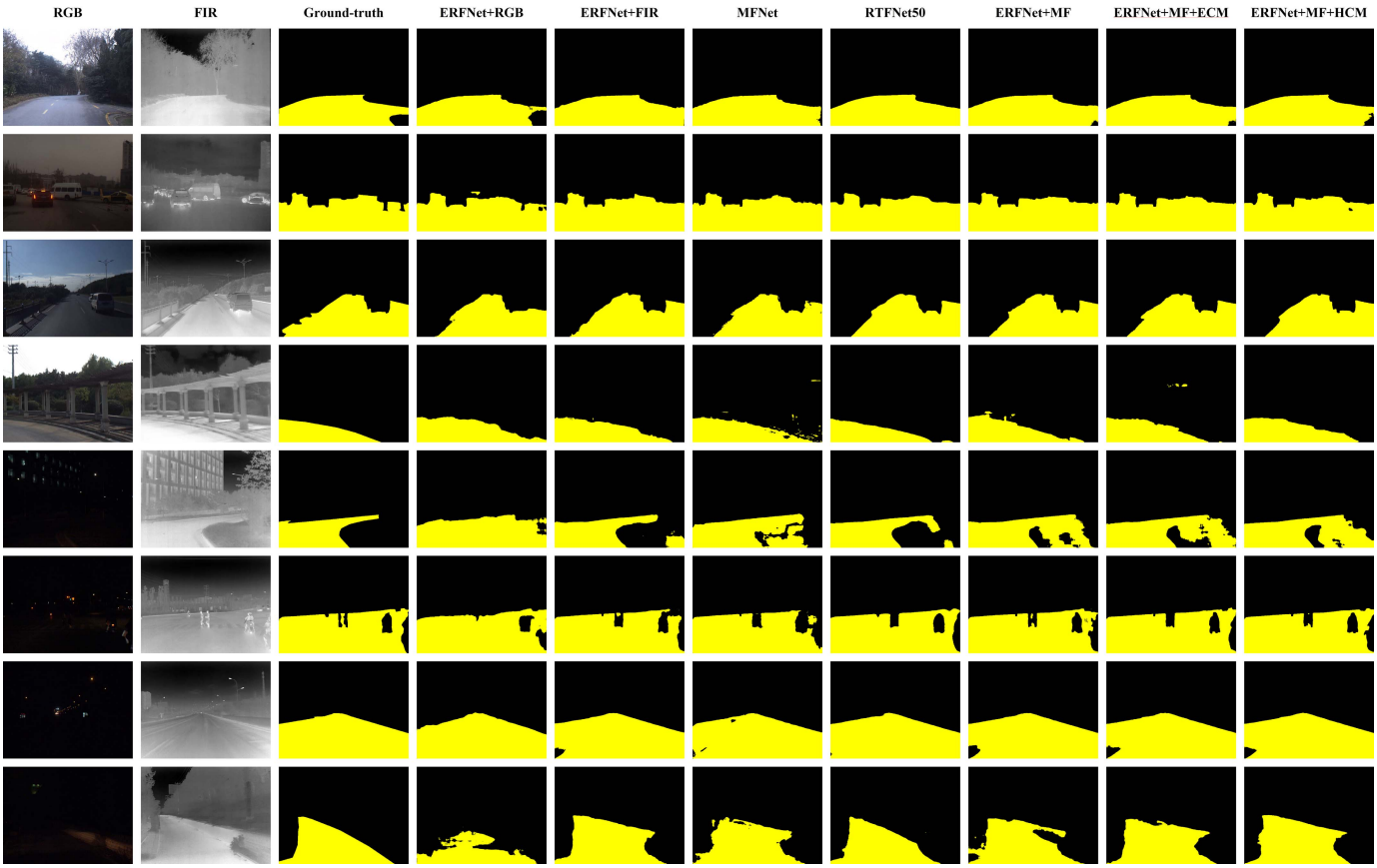



Fig. 9. Some negative results.

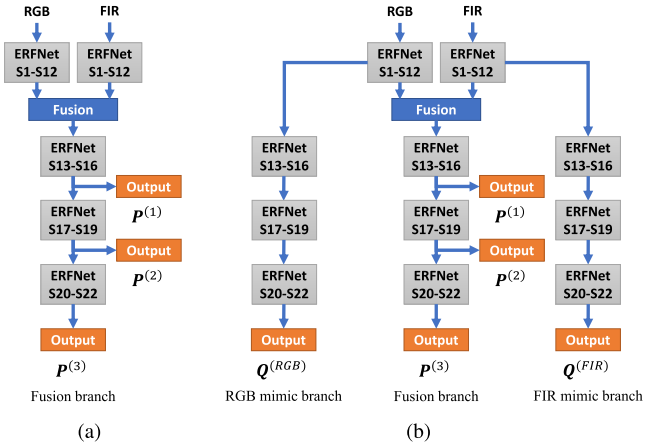



Fig. 10. Models with different fusion branches: (a) ERFNet+MF+HL, (b) ERFNet+MF+ECM+HL.

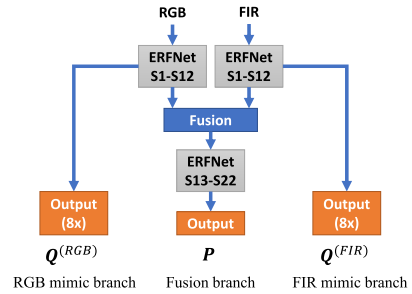



Fig. 11. The ERFNet+MF+ECM(8x) model.

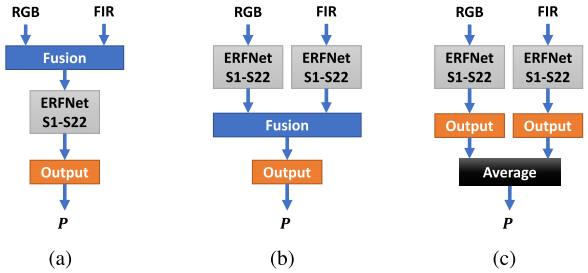



Fig. 12. Concatenation-based fusion models: (a) Early-fusion based ERFNet (ERFNet+EF), (b) Late-fusion based ERFNet (ERFNet+LF), (c) Result-fusion based ERFNet (ERFNet+RF).

that of an extended cross model with general mimic branch. This is because a direct 8x upsampling will lose more detail information when the feature maps with the size of 1/8 is enlarged to the original size.

*3) Hyper-Parameter $\lambda_c$:* Another factor to influence the training effect is the hyper-parameter $\lambda_c$, which regularizes the original segmentation result and the effectiveness of mimic branch. Table VI displays the evaluation results for the ERFNet+MF+HCM when $\lambda_c$ is set to 0.01, 0.1 and 1.0, respectively. Obviously, when $\lambda_c = 0.1$, the ERFNet+MF+HCM gets the best IoU of 97.25% and F-score of 98.61%.

*4) Concatenation Stage:* In this work, the middle-fusion based ERFNet is chosen to construct its ECM and HCM.

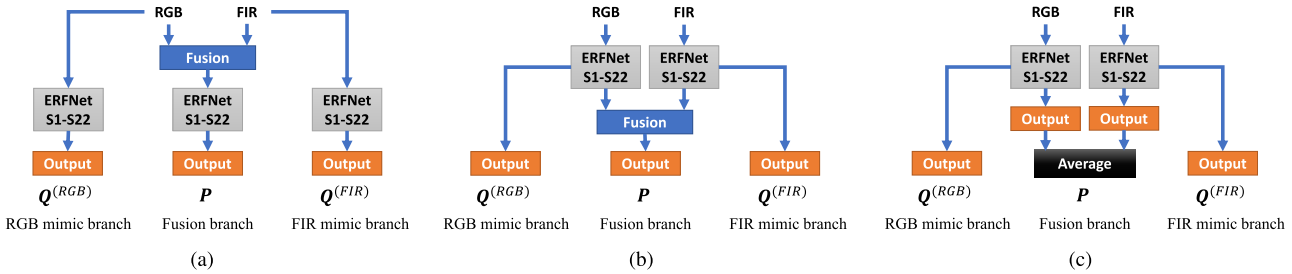



Fig. 13. The extended cross models with different concatenation stages: (a) Early-fusion based ECM (ERFNet+EF+ECM), (b) Late-fusion based ECM (ERFNet+LF+ECM), (c) Result-fusion based ECM (ERFNet+RF+ECM).

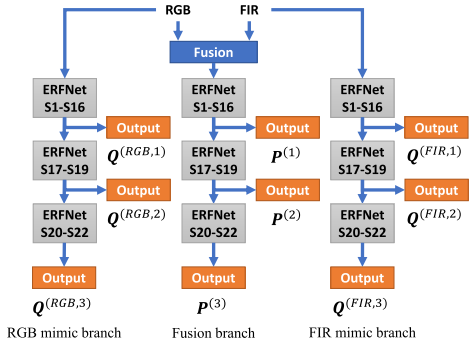



Fig. 14. The hierarchical cross model for early fusion (ERFNet+EF+HCM).

TABLE VI
COMPARISONS ON HYPER-PARAMETER $\lambda_c$ FOR ERFNET+MF+HCM

| $\lambda_c$ | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| 0.01 | **98.28%** | 98.81% | 98.55% | 97.14% |
| 0.1 | **98.28%** | **98.93%** | **98.61%** | **97.25%** |
| 1.0 | 98.25% | 98.87% | 98.56% | 97.16% |

TABLE VII
COMPARISONS ON CONCATENATION STAGES

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| ERFNet+RGB | 96.35% | 97.30% | 96.82% | 93.84% |
| ERFNet+FIR | 97.06% | 98.21% | 97.63% | 95.37% |
| ERFNet+EF | 97.46% | 98.80% | 98.13% | 96.32% |
| ERFNet+MF | 97.88% | **98.96%** | 98.42% | 96.88% |
| ERFNet+LF | 97.92% | 98.59% | 98.25% | 96.56% |
| ERFNet+RF | 97.25% | 98.64% | 97.94% | 95.96% |
| ERFNet+EF+ECM | 98.08% | 98.31% | 98.19% | 96.45% |
| ERFNet+MF+ECM | 98.19% | 98.90% | 98.54% | 97.13% |
| ERFNet+LF+ECM | 98.10% | 98.66% | 98.38% | 96.81% |
| ERFNet+RF+ECM | 97.60% | 98.61% | 98.10% | 96.28% |
| ERFNet+EF+HCM | 97.95% | 98.73% | 98.34% | 96.74% |
| ERFNet+MF+HCM | **98.28%** | 98.93% | **98.61%** | **97.25%** |

Actually, the concatenation stage is also an important factor to affect the fusion result. Consequently, we just build different kinds of fusion models at different concatenation stages. Figure 12 provides three kinds of methods, i.e., the early fusion (ERFNet+EF), late fusion (ERFNet+LF) and result fusion (ERFNet+RF). Table VII presents their performance. We can find that all the metrics of fusion methods are larger than those of the ERFNet+RGB and ERFNet+FIR. This proves the effectiveness of RGB-FIR fusion. Furthermore, among these four fusion methods, the ERFNet+MF obtains the largest IoU of 96.88% and the best F-score of 98.42%. It implies that compared with the other three fusion stages, the middle fusion will fuse the RGB and FIR information more thoroughly.

We also compare the effectiveness of the ECM models corresponding to these three kinds of fusion methods, which is shown in Figure 13, i.e. the ERFNet+EF+ECM, ERFNet+LF+ECM and ERFNet+RF+ECM. Specially, the structure of the ERFNet+RF+ECM is somewhat similar to the original cross model for 3D segmentation in [13], while the structure of the ERFNet+LF+ECM is a bit like the fusion version in [13]. The evaluation results are also reported in Table VII. All these ECM-based approaches obtain better values of IoU and F-score than those without the ECM, which demonstrates the effectiveness of ECM. In addition, the ERFNet+MF+ECM outperforms the other three ECM-based methods on the IoU with 97.13%, which is obviously higher than those of early fusion (96.45%), late fusion (96.81%) and result fusion (96.28%). It implies that the middle fusion is also able to achieve superior results for the ECM-based model.

In view of the architecture of the hierarchical loss, the middle-fusion based HCM, ERFNet+MF+HCM, only compares with the early-fusion one, ERFNet+EF+HCM, illustrated in Figure 14. As presented in Table VII, the ERFNet+EF+HCM is also better than the ERFNet+EF+ECM, which validates its effectiveness. Moreover, like the comparisons on concatenation-based models and their corresponding ECM models, the middle-fusion based method still acquires a higher evaluation result with the IoU of 97.25%, which is much larger than that of the early-fusion based method with the IoU of 96.74%.

### *F. Choice of Base Model*

In this work, the ERFNet is chosen as the base model. There are three reasons. First, as listed in Table III, the ERFNet has fewer parameters of 2.063 million, and the FLOPs of ERFNet are also low, just 17.271 billion. Even after adding the HCM model, the FLOPs are only 43.243 billion, much less than the FLOPs of RTFNet50. This makes the network easier to train in a fast speed, especially when GPU resources are limited. Second, the run-time frame rate of ERFNet is very high. The original ERFNet runs in about 63 frames per second in our experimental environment. Without the mimic branches and auxiliary outputs, the trained ERFNet+MF+HCM model could also run in about 41 frames per second. This satisfies the realtime requirement for autonomous driving. Third, with

TABLE VIII
EVALUATION ON ECM AND HCM MODELS OF PSPNET50

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| PSPNet50+RGB | 96.85% | 96.37% | 96.61% | 93.43% |
| PSPNet50+FIR | 98.05% | 97.15% | 97.60% | 95.31% |
| PSPNet50+Fusion | 98.77% | 97.98% | 98.37% | 96.80% |
| PSPNet50+Fusion+ECM | **98.86%** | 98.03% | 98.44% | 96.93% |
| PSPNet50+Fusion+HCM | 98.83% | **98.10%** | **98.46%** | **96.98%** |

TABLE IX
EVALUATION ON ECM AND HCM MODELS OF SWIFTNETRN-18

| Method | PRE | REC | F-score | IoU |
|---|---|---|---|---|
| SwiftNetRN-18+RGB | 97.20% | 95.82% | 96.51% | 93.25% |
| SwiftNetRN-18+FIR | 96.98% | 97.65% | 97.31% | 94.77% |
| SwiftNetRN-18+Fusion | **98.52%** | 97.96% | 98.24% | 96.54% |
| SwiftNetRN-18+Fusion+ECM | 98.44% | 98.21% | 98.33% | 96.71% |
| SwiftNetRN-18+Fusion+HCM | 98.48% | **98.22%** | **98.35%** | **96.75%** |

fewer parameters and FLOPs, the ERFNet still achieves relatively good performance.

In fact, besides the ERFNet, our ECM and HCM models are also suitable for other networks. Here we show the results when the PSPNet50 [64] and SwiftNetRN-18 [65] are regarded as the base model, respectively.

The PSPNet50 [64] is based on the backbone ResNet50 [58]. Then a pyramid pooling module (PPM) is utilized after the conv5_x block to obtain sub-region representations at different scales, followed by upsampling and concatenation to generate feature maps with both global and local context information. At last these feature maps are processed by a convolution layer to predict the segmentation results. To construct the fusion network "PSPNet50+Fusion", we add the fusion block after the conv3_x block in the ResNet50. Furthermore, based on the fusion model "PSPNet50+Fusion", the output branches for RGB and FIR feature maps after fusion block are also preserved, which constitutes the extended cross model "PSPNet50+Fusion+ECM". Additionally, the auxiliary loss after the conv4_x block in the ResNet50 is reserved for the auxiliary loss in the hierarchical cross model "PSPNet50+Fusion+HCM".

The quantitative evaluation results for PSPNet50 on the whole testing set are shown in Table VIII. Like the results from the ERFNet, the PSPNet50+RGB still obtains higher scores, which indicates that the PSPNet50 could also process RGB images in the nighttime. The F-measure and IoU scores of the PSPNet50+Fusion are better than those of the PSPNet50+RGB and PSPNet50+FIR, which illustrates that the fusion of RGB and FIR can also improve the performance of the PSPNet50. Moreover, our PSPNet50+Fusion+HCM and PSPNet50+Fusion+ECM models achieve the best and the second-best performance, respectively. This demonstrates the superiority of our ECM and HCM models.

The SwiftNetRN-18 [65] is based on the backbone ResNet18 [58]. It follows an encoder-decoder structure, where the generated low-resolution feature maps from the backbone encoder are upsampled to the original scale progressively. Besides, the skip connections connect between the encoder and decoder parts to provide more details for semantic segmentation. To construct the fusion network "SwiftNetRN-18+Fusion", we add the fusion block after the spatial pyramid pooling (SPP) module with the backbone ResNet18, while the skip connections in the original SwiftNetRN-18 model are all retained and connect to the corresponding layers in the fusion branch. Next, based on the fusion model "SwiftNetRN-18+Fusion", the output branches for RGB and FIR feature maps after fusion block are also preserved, building the extended cross model "SwiftNetRN-18+Fusion +ECM". Here the skip connections from the encoder of RGB and FIR branch also connect to their corresponding layers in their respective decoder. At last, to construct the hierarchical cross model "SwiftNetRN-18+Fusion+HCM", the auxiliary losses are added after each upsampling block in the decoder.

The quantitative evaluation results for the SwiftNetRN-18 on the whole testing subset are listed in Table IX. Similarly, the fusion-based models, the SwiftNetRN-18+Fusion, SwiftNetRN-18+Fusion+ECM and SwiftNetRN-18+Fusion+HCM, all obtain better F-measure and IoU scores on the whole testing subset, which indicates that fusion of RGB and FIR is able to improve the performance of the SwiftNetRN-18. In addition, the HCM and ECM based models, SwiftNetRN-18+Fusion+HCM and SwiftNetRN-18+Fusion+ECM still achieve the best and the second-best performance, respectively. This also demonstrates the superiority of our ECM and HCM models once again.

## VI. CONCLUSION

In this paper, we first construct an RGB-FIR dataset by a hybrid camera for road detection, where the RGB and FIR images are aligned pixel-wisely, and the ghost phenomenon caused by thermal camera rolling-shutter effect is attenuated. In view of the concatenation-based ERFNet model, we design an extended cross model, which restores the removed layers of RGB and FIR branches after the fusion block to imitate the probability distributions of the fusion outputs by using KL-divergence. Moreover, the hierarchical loss structure is introduced to build a hierarchical cross model for better performance. The experiments on our RGB-FIR road dataset demonstrate the superiority and effectiveness of our proposed middle-fusion based extended cross model and hierarchical cross model.

## REFERENCES

[1] S. Hwang, Y. Choi, N. Kim, K. Park, J. S. Yoon, and I. S. Kweon, "Low-cost synchronization for multispectral cameras," in *Proc. Int. Conf. Ubiquitous Robot. Ambient Intell.*, 2015, pp. 435–436.

[2] Y. Zhang *et al.*, "Build your own hybrid thermal/EO camera for autonomous vehicle," in *Proc. IEEE Conf. Robot. Autom.*, May 2019, pp. 6555–6560.

[3] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1037–1045.

[4] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.

[5] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *CoRR*, vol. abs/1803.05347, pp. 1–14, Aug. 2019.

[6] M. Ye, Z. Wang, X. Lan, and P. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. IJCAI*, 2018, p. 2.

[7] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Dec. 2019, pp. 3622–3631.

[8] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proc. AAAI*, 2020, pp. 11345–11353.

[9] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2018, pp. 1–5.

[10] A. Kantarci and H. K. Ekenel, "Thermal to visible face recognition using deep autoencoders," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2019, pp. 1–5.

[11] Y. Zhu, C. Li, Y. Lu, L. Lin, B. Luo, and J. Tang, "FANet: Quality-aware feature aggregation network for RGB-T tracking," 2018, *arXiv:1811.09855*.

[12] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for RGB-infrared object tracking," *Pattern Recognit. Lett.*, vol. 130, pp. 12–20, Feb. 2020.

[13] M. Jaritz, T.-H. Vu, R. de Charette, E. Wirbel, and P. Perez, "XMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12605–12614.

[14] H. Kong, J. Audibert, and J. Ponce, "Vanishing point detection for road detection," in *Proc. CVPR*, Jun. 2009, pp. 96–103.

[15] C. Rasmussen, "Grouping dominant orientations for ill-structured road following," in *Proc. IEEE Comput. SoC. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2004, pp. 1–5.

[16] K. Lu, J. Li, X. An, and H. He, "A hierarchical approach for road detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Dec. 2014, pp. 517–522.

[17] J. W. Lee and U. K. Yi, "A lane-departure identification based on LBPE, Hough transform, and linear regression," *Comput. Vis. Image Understand.*, vol. 99, no. 3, pp. 359–383, 2005.

[18] J. Fritsch, T. Kuhnl, and F. Kummert, "Monocular road terrain detection by combining visual and spatial information," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1586–1596, Aug. 2014.

[19] C. Tan, T. Hong, T. Chang, and M. Shneier, "Color model-based real-time learning for road following," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Dec. 2006, pp. 939–944.

[20] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 309–318, Dec. 2004.

[21] O. Y. Agunbiade, T. Zuva, A. O. Johnson, and K. Zuva, "Enhancement performance of road recognition system of autonomous robots in shadow scenario," 2014, *arXiv:1401.2051*.

[22] Z. Ying, G. Li, X. Zang, R. Wang, and W. Wang, "A novel shadow-free feature extractor for real-time road detection," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 611–615.

[23] Z. Wang, G. Cheng, and J. Zheng, "Road edge detection in all weather and illumination via driving video mining," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 232–243, Jun. 2019.

[24] J. M. Álvarez, T. Gevers, Y. LeCun, and A. M. López, "Road scene segmentation from a single image," in *Proc. ECCV*, 2012, pp. 376–389.

[25] R. Mohan, "Deep deconvolutional networks for scene parsing," 2014, *arXiv:1411.4101*.

[26] J. Mu noz-Bulnes, C. Lopez, I. Parra, D. Llorca, and M. Sotelo, "Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Dec. 2017, pp. 366–371.

[27] M. Teichmann, M. Weber, J. M. Zöllner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Feb. 2018, pp. 1013–1020.

[28] Z. Chen and Z. Chen, "RBNet: A deep neural network for unified road and road boundary detection," in *Proc. ICONIP*, 2017, pp. 677–687.

[29] X. Han, J. Lu, C. Zhao, S. You, and H. Li, "Semisupervised and weakly supervised road detection based on generative adversarial networks," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 551–555, Apr. 2018.

[30] A. Asvadi, C. Premebida, P. Peixoto, and U. Nunes, "3D LIDAR-based static and moving obstacle detection in driving environments: An approach based on voxels and multi-region ground planes," *Robot. Auto. Syst.*, vol. 83, pp. 299–311, Sep. 2016.

[31] X. Hu, F. S. A. Rodriguez, and A. Gepperth, "A multi-modal system for road detection and segmentation," in *Proc. IEEE Intell. Veh. Symp. Process.*, Jun. 2014, pp. 1365–1370.

[32] L. Chen, J. Yang, and H. Kong, "Lidar-histogram for fast road and obstacle detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Dec. 2017, pp. 1343–1348.

[33] Y. Zhang, S. Gu, J. Yang, J. Álvarez, and H. Kong, "Fusion of LiDAR and camera by scanning in LiDAR imagery and image-guided diffusion for urban road detection," in *Proc. IEEE Intell. Veh. Symp. (IV)*, May 2018, pp. 579–584.

[34] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893.

[35] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4376–4382.

[36] C. Xu *et al.*, "SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation," 2020, *arXiv:2004.01803*.

[37] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast LIDAR-based road detection using fully convolutional neural networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1019–1024.

[38] L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, "CRF based road detection with multi-sensor fusion," in *Proc. IV*, Jun. 2015, pp. 192–198.

[39] L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, "Hybrid conditional random field based camera-LiDAR fusion for road detection," *Inf. Sci.*, vol. 432, pp. 543–558, Mar. 2018.

[40] S. Gu, Y. Zhang, J. Tang, J. Yang, J. M. Alvarez, and H. Kong, "Integrating dense LiDAR-camera road detection maps by a multi-modal CRF model," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11635–11645, Dec. 2019.

[41] Z. Chen, J. Zhang, and D. Tao, "Progressive LiDAR adaptation for road detection," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 693–702, May 2019.

[42] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. ECCV*, 2020, pp. 340–356.

[43] H. Wang, R. Fan, P. Cai, and M. Liu, "SNE-RoadSeg+: Rethinking depth-normal translation and deep supervision for freespace detection," 2021, *arXiv:2107.14599*.

[44] X. Lv, Z. Liu, J. Xin, and N. Zheng, "A novel approach for detecting road based on two-stream fusion fully convolutional network," in *Proc. IEEE Intell. Veh. Symp. (IV)*, May 2018, pp. 1464–1469.

[45] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view LiDAR point cloud and front view camera image for 3D object detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1–6.

[46] F. Yang, H. Wang, and Z. Jin, "A fusion network for road detection via spatial propagation and spatial transformation," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107141.

[47] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. ACCV*, 2016, pp. 212–228.

[48] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robot. Autom. Lett.*, vol. 5, pp. 5558–5565, 2020.

[49] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.

[50] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8599–8608.

[51] K. Xiang, K. Yang, and K. Wang, "Polarization-driven semantic segmentation via efficient attention-bridged fusion," *Opt. Exp.*, vol. 29, 4, pp. 4802–4820, 2021.

[52] J. Zhang, K. Yang, and R. Stiefelhagen, "ISSAFE: Improving semantic segmentation in accidents by fusing event-based data," 2020, *arXiv:2008.08974*.

[53] J. S. Yoon *et al.*, "Thermal-infrared based drivable region detection," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Dec. 2016, pp. 978–985.

[54] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

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

[56] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[57] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–15, Oct. 2015.
[59] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 243–250.
[60] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.
[61] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4885–4891.
[62] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[63] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[64] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2017, pp. 6230–6239.
[65] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12599–12608.

**Yigong Zhang** is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology, Nanjing, China. His research interests include computer vision and sensor fusion.

**Jin Xie** received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include image forensics, computer vision, and machine learning. Currently, he is focusing on 3-D computer vision with the convex optimization and deep learning methods.

**José M. Álvarez** received the Ph.D. degree from the Autonomous University of Barcelona (UAB), Bellaterra, Spain, in 2010. During the Ph.D. program, he visited the ISLA Group, University of Amsterdam, from 2008 to 2009, and the Group Research Electronics, Volkswagen, Germany, in 2010. Subsequently, he was a Post-Doctoral Researcher with the Courant Institute of Mathematical Science, New York University. He was a Computer Vision Researcher with CSIRO-Data61, Australia, from 2016 to 2018. He is currently a Senior Deep Learning Research Scientist with NVIDIA, Santa Clara, CA, USA. He received the Best Ph.D. Thesis Award from UAB in 2010. Since 2014, he has been serving as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

**Cheng-Zhong Xu** (Fellow, IEEE) received the Ph.D. degree from The University of Hong Kong in 1993. He is currently the Chair Professor of the Department of Computer and Information Science and the Dean of the Faculty of Science and Technology, University of Macau, China, and the Director of the Institute of Advanced Computing and Data Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include parallel and distributed systems and cloud computing. He serves on a number of journal editorial boards, including IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON CLOUD COMPUTING, *Journal of Parallel and Distributed Computing*, and *Science China Information Sciences*.

**Jian Yang** received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NJUST) in 2002. In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology. He is currently a Changjiang Distinguished Professor with the School of Computer Science and Technology, NJUST. His research interests include pattern recognition, computer vision, and machine learning. He is also an Associate Editor of *Pattern Recognition Letters* and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

**Hui Kong** received the Ph.D. degree in computer vision from Nanyang Technological University, Singapore, in 2007. He is currently an Associate Professor with the State Key Laboratory of Internet of Things for Smart City (SKL-IOTSC), Department of Electromechanical Engineering (EME), University of Macau, Macau. His research interests include sensing and perception for autonomous driving, SLAM, mobile robotics, multi-view geometry, and motion planning.