GMNet: Graded-Feature Multilabel-Learning Network for RGB-Thermal Urban Scene Semantic Segmentation

Wujie Zhou, *Member, IEEE*, Jinfu Liu, Jingsheng Lei, Lu Yu, *Member, IEEE*, and Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*— Semantic segmentation is a fundamental task in computer vision, and it has various applications in fields such as robotic sensing, video surveillance, and autonomous driving. A major research topic in urban road semantic segmentation is the proper integration and use of cross-modal information for fusion. Here, we attempt to leverage inherent multimodal information and acquire graded features to develop a novel multilabel-learning network for RGB-thermal urban scene semantic segmentation. Specifically, we propose a strategy for graded-feature extraction to split multilevel features into junior, intermediate, and senior levels. Then, we integrate RGB and thermal modalities with two distinct fusion modules, namely a shallow feature fusion module and deep feature fusion module for junior and senior features. Finally, we use multilabel supervision to optimize the network in terms of semantic, binary, and boundary characteristics. Experimental results confirm that the proposed architecture, the graded-feature multilabel-learning network, outperforms state-of-the-art methods for urban scene semantic segmentation, and it can be generalized to depth data.**

***Index Terms*— RGB-thermal semantic segmentation, graded-features, cross-modal fusion, multilabel-learning, refinement strategy.**

I. Introduction

SEMANTIC segmentation refers to the dense labeling of each pixel on an image into a class. In recent years, semantic segmentation has shifted from traditional computer vision algorithms to deep learning methods, with convolutional neural networks (CNNs) demonstrating their effectiveness and reliability [1]–[4]. Most deep learning-based semantic segmentation methods are designed using multimodal sensory data, such as RGB images acquired by conventional cameras and depth maps acquired by depth sensors. Depth maps are highly effective to complement RGB images, which may lose information under adverse lighting conditions, such as dimness or darkness. However, depth maps may also provide unsatisfactory results in some cases. For instance, the time of flight, the main measurement principle of most depth cameras, is susceptible to multiple reflections. In addition, depth maps may be affected by blur under high-speed sensor/scene motions as those experienced during autonomous vehicle driving. Alternatively, we found that thermal imaging sensors provide robust imaging under poor lighting conditions. These images reflect the thermal radiation of any object with a temperature above the absolute zero and are acquired by thermal imaging sensors [5]. Thus, thermal images can overcome the limitation of depth maps for applications such as autonomous driving and a variety of challenging conditions. Therefore, we adopt thermal images to complement RGB images under adverse lighting conditions to achieve semantic segmentation.

The multimodal multilevel feature fusion strategy [6] is essential in many computer vision tasks, such as semantic segmentation [7], co-attention [8], object detection [9], and classification [10]. Thus, we leverage this concept for image semantic segmentation based on RGB and thermal cues. To efficiently fuse these cues, several multimodal fusion strategies have been developed [11]–[13], achieving promising results. However, existing RGB-T (RGB-thermal) semantic segmentation methods [14]–[16] still face two main challenges: 1) fully leveraging multilevel features and 2) effectively integrating cross-modal RGB and thermal features.

Regarding multilevel features, "senior" features provide discriminative semantic information as guidance for labeling each pixel during segmentation, while "junior" features provide details such as spatial and texture information for accurate segmentation, especially along edges [17]. Existing RGB-T semantic segmentation methods leverage multilevel features via progressive feature merging [16] and generate a refined feature as the final prediction. This approach fails to fully use level-specific characteristics and directly fuse multilevel features. Consequently, low-level features may

Manuscript received February 9, 2021; revised June 23, 2021 and August 3, 2021; accepted August 27, 2021. Date of publication September 8, 2021; date of current version September 14, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61502429 and Grant 61972357 and in part by Zhejiang Provincial Natural Science Foundation of China under Grant LY18F020012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Raymond Fu. *(Corresponding author: Wujie Zhou.)*

Wujie Zhou is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China, and also with the Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: wujiezhou@163.com).

Jinfu Liu and Jingsheng Lei are with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China.

Lu Yu is with the Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China.

Jenq-Neng Hwang is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98105 USA.

Digital Object Identifier 10.1109/TIP.2021.3109518

introduce noise [18], [19], which can, in turn, affect the semantic segmentation accuracy.

Existing methods combine cross-modal features by considering the depth map or thermal image as a fourth information channel or fusing it with the RGB image by simple summation [15], [16]. These methods treat RGB and thermal (or depth) data as having the same information and neglect the fact that images from other modalities are generally more informative than RGB images under adverse lighting conditions. Moreover, RGB data can provide rich color and texture information under appropriate lighting conditions. Thus, simply adding these information modalities is ineffective because their properties are lost. However, most semantic segmentation methods use a unified fusion strategy across levels. Furthermore, senior features typically carry more semantic information than junior features that generally provide details and edge information. Thus, a unified fusion strategy may be inappropriate for deep networks that perform semantic segmentation.

To address the abovementioned problems, we propose the graded-feature multilabel-learning network (GMNet) for RGB-T urban scene semantic segmentation.

The main contributions of this work are:

1) We exploit multilevel features and divide them into three grades: "senior" features with abundant semantic cues, "intermediate" features, and "junior" features with abundant details and edge cues. We also suppress distractors at the lower layers by using a refinement strategy. Furthermore, three supervision strategies are simultaneously applied to the graded features.

2) To fully integrate cross-modal informative cues, we introduce two distinct fusion modules, the shallow feature fusion module (SFFM) and the deep feature fusion module (DFFM) that are applied according to specific characteristics at different levels.

3) Semantic, binary, and boundary loss functions are used to optimize the parameters of the proposed GMNet. Experimental results confirm that the proposed GMNet considerably outperforms state-of-the-art (SOTA) methods on two RGB-T datasets. Another evaluation on the RGB-D dataset demonstrates the ability of GMNet to generalize to RGB-D data.

The remainder of this paper is organized as follows. Section II reviews related studies. Section III details the proposed GMNet. Section IV presents the experimental results and discussion. Finally, Section V presents conclusions and directions of future study.

II. Related Work

A. Single-Modal Semantic Segmentation Methods

Shelhamer *et al*. [20] first proposed the fully convolutional network to perform end-to-end semantic segmentation. They designed a fully convolutional network for pixel-level image segmentation by amending image classification networks (e.g., VGG-16 [21]). Then, Noh *et al*. [22] designed DeconvNet by employing a convolutional module to extract features and a deconvolutional module to restore the resolution. Badrinarayanan *et al*. [23] developed SegNet by employing an encoder–decoder framework. The decoder and encoder played similar roles to those of the deconvolutional and convolutional networks in DeconvNet. To preserve spatial information and leverage low-level features, Ronneberger *et al*. [24] proposed the UNet by designing skip networks between the decoder and encoder. Despite UNet being originally designed for biomedical imaging, it can be suitably extended to other fields. Paszke *et al.* [25] developed ENet for real-time image understanding. Zhao *et al*. [26] acknowledged the benefits of contextual information for semantic segmentation. Accordingly, they proposed PSPNet using a pyramid pooling module to extract local and global contextual cues at disparate scales. Rather than using bilinear upsampling to restore resolution, Wang *et al.* [27] employed the DUCNet to decode and maintain detailed information. In addition, the hybrid dilated convolution was introduced for feature extraction. Yu *et al*. [28] proposed DFNet to capture multiscale contextual information and employed deep semantic boundary supervision to distinguish boundary features. Yu *et al*. [29] proposed BiSeNet to preserve spatial information with a spatial path and acquired sufficient receptive field information with a contextual path. Sun *et al*. [30] proposed HRNet to preserve high-resolution information throughout feature extraction. Subsequently, they modified HRNet by integrating upsampled representations from parallel convolutions, rather than using representations from the high-resolution convolution [31] as in the original HRNet. Kong and Fowlkes [32] proposed a pixelwise attention gating unit for scene parsing. The unit can learn to select the appropriate spatially varying pooling and dynamically allocate computations to different image regions. Other relevant studies can be found in [33], [34]. Although single-modal semantic segmentation has been thoroughly studied and has achieved promising results in recent years, most methods are unsatisfactory under complicated scenes, including those with cluttered backgrounds [35] and adverse lighting conditions (e.g., dimness and darkness).

B. RGB-D Semantic Segmentation Networks

Depth information can complement RGB cues by providing rich spatial information about distances [36] and can substantially improve the segmentation of challenging scenes. Hazirbas *et al*. [37] proposed FuseNet by employing an encoder–decoder structure to integrate RGB and depth information. They fused depth feature maps into RGB feature maps as the network proceeded deeper across its layers. Wang and Neumann [38] suitably handled geometric information in traditional CNNs by using depth-aware average pooling and depth-aware convolution operations. In addition, the depth similarities between pixels were leveraged for incorporating geometric information. Lee *et al*. [39] proposed RDFNet to extract and integrate multilevel RGB-depth (RGB-D) information by applying the concept of residual learning to RGB-D semantic segmentation. Gao *et al.* [40] introduced a global CNN based on the large kernel concept to improve the performance of a dual-encoder fusion network. Jiang *et al.* [41] proposed RedNet with a residual module and a skip connection

between the decoder and encoder that bypasses spatial features. Zhou *et al.* [42] proposed a three-branch self-attention network for indoor RGB-D semantic segmentation. In addition to the RGB and depth branches, a cross-modal distillation branch with a self-attention module extracts joint RGB-D features at each level. Deng *et al*. [43] built a RFBNet by using a residual fusion block that reflects interdependencies between modality-specific streams. Hu *et al*. [12] proposed ACNet to integrate RGB and depth cues in a proportion determined by the input. Furthermore, a dedicated ACNet stream, in addition to the streams for RGB and depth data, processes the fusion features. Chen *et al.* [44] proposed a bidirectional cross-modality-guided encoder for RGB-D semantic segmentation. Separation-and-aggregation gating performs filtering and recalibration of the dual-stream features, and bidirectional multistep propagation preserves the specificities of the two streams during information fusion. Lin *et al.* [45] proposed a switchable context network by jointly using RGB and depth data to guide feature aggregation. Xiong *et al*. [46] leveraged a variational context-deformable network for the structured learning of adaptive receptive fields. The network learns a deformable spatial context under the guidance of depth data. Zhang *et al*. [47] designed NANet to aggregate non-local RGB-D information along the spatial and channel dimensions. The abovementioned methods have confirmed that depth information can greatly improve segmentation. However, depth data may be noisy and lead to unsatisfactory results in some cases.

C. RGB-T Semantic Segmentation Methods

Recently, many studies have attempted to employ other data sources to complement RGB information, obtaining a high performance. MFNet [14], RTFNet [15], PSTNet [48], and FuseSeg [16] have been proposed to leverage RGB-T information. Ha *et al*. [14] proposed MFNet by integrating thermal and RGB features. They extracted RGB-T information from two symmetric encoders and introduced a mini-inception module in the encoder. Sun *et al.* [15] designed RTFNet that includes two encoders and one decoder. In the encoder, the thermal features are gradually integrated into the RGB cues. In the decoder, two types of upsampling inception blocks extract features and restore their resolution. Shivakumar *et al*. [48] designed PSTNet with a dual-stream CNN to reuse the RGB stream and fuse RGB and thermal information for real-time inference. Sun *et al*. [16] employed DenseNet structures [49] as the encoders in their FuseSeg. In addition, they proposed a two-stage fusion strategy applied in the encoder and decoder. However, these methods mostly use simple and unitary fusion strategies, which may lead to insufficient integration and the underuse of specific characteristics of multilevel features.

III. Proposed GMNet

A. Architecture

Fig. 1 shows the proposed GMNet that includes two symmetric encoders for feature extraction and three grading decoding stages for original resolution restoration. The two symmetric encoders are based on ResNet [50] as the backbone and consider RGB and thermal images as their respective inputs. We introduced the division of multilevel features into three types of graded features: senior, intermediate, and junior. Considering the last three layers in ResNet, compared with the first two layers, consist of more residual blocks containing the downsampling operations, which effectively enlarge the visual receptive fields and improve the representation of semantic feature, we selected the features extracted from these layers as our senior features. Correspondingly, the features from the first layer, owing to their advantage of retaining more detailed information, are selected as our junior features. In addition, for balance, the features from the two layers in the middle are selected as our intermediate features. Next, three grading decoding stages were included in the GMNet. In the first decoding stage, the features extracted from the last three layers (i.e., the senior features) were integrated via the proposed DFFM and further aggregated by a stepwise decoder to generate a senior semantic map S_1. In the second decoding stage, the features extracted from the two layers in the middle (i.e., the intermediate features) were refined by elementwise multiplication with the senior semantic map S_1 to then obtain an intermediate semantic map S_2. In the third decoding stage, the features extracted from the first layer (i.e., the junior features) were similarly refined by the semantic map S_1 to obtain a junior semantic map S_3. We adopted SFFM for fusion in the last two decoding stages, which is simpler than the fusion in the first decoding stage.

B. Encoding

We designed the RGB and thermal encoders with symmetric structures, except for the input channel size in the first layer. Unlike MFNet [14] and FuseSeg [16], we used ResNet [50] as the feature extractor. As the fully connected and average pooling layers of ResNet may lead to the loss of spatial information and details, we removed them for feature extraction. ResNet has an initial module that consists of a convolutional layer, followed by a batch normalization (BN) layer, and a rectified linear unit (ReLU) activation layer. We modified the input channel size of the convolutional layer in the initial module of the thermal encoder to one because ResNet is intended for three-channel RGB images. After the initial module, a max pooling layer and four subsequent residual layers gradually increase the channel size of the feature maps and reduce the resolution. More details about ResNet and its residual layers can be found in [50].

C. Grading Decoding Stages

The three proposed grading decoder streams enabled label decoupling (see Fig. 1). With label decoupling, the semantic label was represented as foreground, background, and boundary maps, which were considered for supervision during model learning. To fully use the extracted multilevel features from the RGB and thermal modalities, we manually split the features into three grades, namely senior, $\mathrm{S} = \{F_3^{fused}, F_4^{fused}, F_5^{fused}\}$, intermediate, $\mathrm{M} = \{F_2^{fused}, F_3^{fused}\}$, and junior, $\mathrm{J} = \{F_1^{\mathrm{fused}}\}$ features. Then, we used rich semantic cues in cross-modal high-level features to refine cross-modal low-level

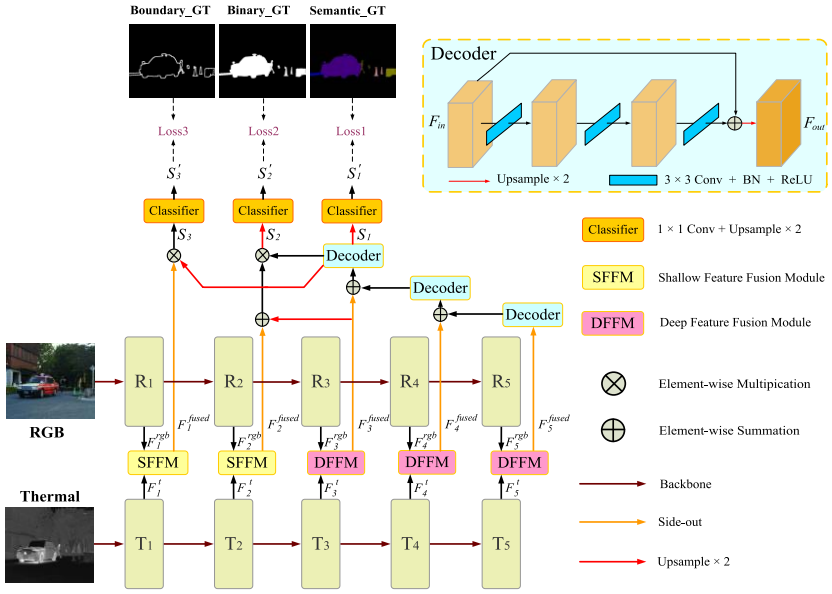

Fig. 1. Overall structure of the proposed GMNet.

features by suppressing background distractors. Specifically, we propose a refinement strategy to leverage the features in the three grades. We first generated cross-modal feature maps $\{F_i^{fused};\ i = 1, 2, \ldots, 5\}$ by integrating RGB and thermal features after the SFFM (generating F_1^{fused} and F_2^{fused}) and the distinct DFFM (generating F_3^{fused}, F_4^{fused}, and F_5^{fused}).

In the first decoding stage, the three cross-modal feature maps (i.e., F_3^{fused}, F_4^{fused}, and F_5^{fused}) were aggregated by iterative decoding to then generate the senior semantic map S_1:

$$S_1 = F_{D3}\left(F_{D2}\left(F_{D1}\left(F_5^{fused}\right) \oplus F_4^{fused}\right) \oplus F_3^{fused}\right), \quad (1)$$

where $\oplus$ represents elementwise summation, and $F_{Di}(i = 1, 2, 3)$denotes a simple decoder formulated as follows:

$$F_{out} = F_{up}\left(Conv\left(F_{in}\right) \oplus F_{in}\right), \quad (2)$$

where F_{out} and F_{in} are the decoder output and input, respectively; *Conv* denotes three cascade convolutional blocks; and F_{up} denotes the upsampling operation.

Subsequently, the senior feature map S_1 contained rich semantic information that guided the other two stages (i.e., grades intermediate M and junior J).

In the second decoding stage, we integrated features F_2^{fused} and F_3^{fused} from grade M by elementwise addition and refined them with the senior semantic map S_1, to obtain the intermediate semantic map S_2:

$$S_2 = \left(F_2^{fused} \oplus \uparrow F_3^{fused}\right) \otimes S_1, \quad (3)$$

where $\otimes$ represents elementwise multiplication, and $\uparrow$ denotes the upsampling operation.

In the third decoding stage, F_1^{fused} was directly used and similarly refined with the senior semantic map S_1 to obtain junior semantic map S_3:

$$S_3 = F_1^{fuesed} \otimes \uparrow S_1. \quad (4)$$

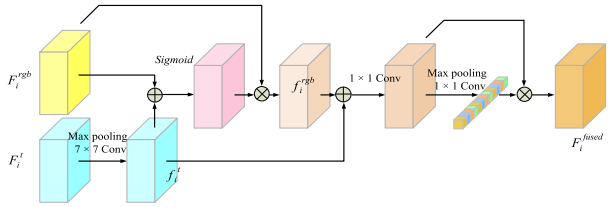

Fig. 2. Structure of the proposed SFFM.

Finally, we used three classifiers to generate three prediction maps for the senior, intermediate, and junior features denoted as S_1', S_2', and S_3', respectively. Each classifier has the same structure with a convolutional layer that changes the number of channels of the feature map and an upsampling layer.

D. Shallow Feature Fusion Module (SFFM)

There are two major problems when attempting to merge RGB and thermal features. One problem is the compatibility between features by the inherent differences in the RGB and thermal modalities. Another problem is that raw RGB images can carry misleading and incomplete information under adverse lighting conditions. Such information may misguide segmentation. Thermal images are robust to various adverse conditions, including poor lighting, and may contribute to the segmentation accuracy.

Considering the abovementioned problems, we propose the SFFM shown in Fig. 2. A global max pooling operation and a 7 × 7-kernel convolution improve the extraction of spatial information in the thermal images as follows:

$$f_i^t = Conv_7\left(P_{\max}\left(F_i^t\right)\right), \quad i = 1, 2, \quad (5)$$

where F_i^t is the feature derived from thermal layer i (we only applied the SFFM to the first two layers), $P_{\max}$ denotes the global max pooling per point in the features along the channel axis, $Conv_7$ denotes a 7 × 7-kernel convolution, and

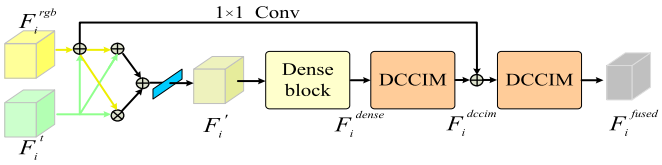

Fig. 3. Proposed DFFM.

f_i^t is the enhanced thermal information used to guide the RGB information.

Then, we refined the information from the RGB stream with the enhanced thermal information as follows:

$$f_i^{rgb} = Sig\left(F_i^{rgb} \oplus f_i^t\right) \otimes F_i^{rgb}, \quad i = 1, 2, \tag{6}$$

where F_i^{rgb} is the feature derived from RGB layer i ($i = 1$, 2), f_i^{rgb} is the refined RGB information, and *Sig* denotes the elementwise sigmoid function.

Then, we fused the refined RGB information and enhanced the thermal information by elementwise summation. In addition, we further improved the fused features via channel-wise attention. Hence, we generated fused features F_i^{fused} in the first two layers as follows:

$$F_i^{fused} = C_{att}\left(Conv_1\left(f_i^{rgb} \oplus f_i^t\right)\right), \quad i = 1, 2, \tag{7}$$

where F_i^{fused} represents the fused features from RGB and thermal layers i ($i = 1$, 2), $Conv_1$ denotes a 1×1-kernel convolution used to change the channel size to 64, and C_{att} denotes the channel-wise attention function defined as

$$C_{att}(f) = T\left(C_{\max}(f)\right) \otimes f, \tag{8}$$

where f is the input features, $C_{\max}$ denotes the global max pooling operation per feature map, and T denotes a two-layer perceptron.

E. Deep Feature Fusion Module (DFFM)

Unlike shallow layer features, deep layer features provide richer contextual semantic information. Thus, we proposed the DFFM to fuse the dual-stream information in the last three layers (see Fig. 3). In the DFFM, we used four steps to integrate and enhance the features.

In the first step, we fused the features from the two streams and obtained raw fused features F_i' ($i = 3$, 4, 5) as follows:

$$F_i' = \left(F_i^{rgb} \oplus F_i^t \oplus F_i^t\right) \oplus \left(F_i^t \otimes \left(F_i^t \oplus F_i^{rgb}\right)\right), \quad i = 3, 4, 5, \tag{9}$$

where F_i' represents the raw fused features, and F_i^{rgb} and F_i^t denote the features from RGB and thermal layers i ($i = 3$, 4, 5), respectively.

Inspired by DenseNet [49] to enhance the transmission of features and prevent the vanishing gradient problem, we passed raw fused features F_i' through the dense block shown in Fig. 4 and generated dense features F_i^{dense} as follows:

$$F_i^{dense} = Conv_3\left(F_i' \oplus Conv_3\left(Cat\left(F_i', D\left(F_i'\right)\right)\right)\right), \quad i = 3, 4, 5, \tag{10}$$

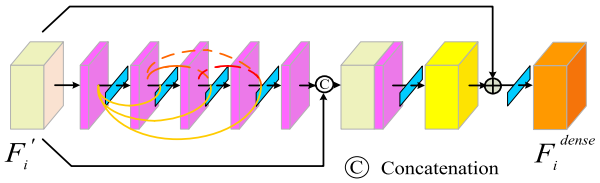

Fig. 4. Dense block in DFFM.

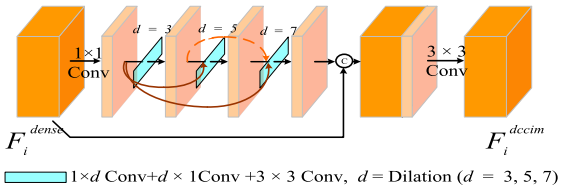

Fig. 5. Proposed DCCIM.

where *Cat* denotes the concatenation of its arguments along the channel axis. Further, $Conv_3$ denotes a 3×3-kernel convolution, where the second convolutional layer changes the channel size of the features to 64 to reduce the number of parameters, whereas the first convolutional layer does not change, and D denotes four densely connected 3×3-kernel convolutions.

Then, we introduced the densely cascaded contextual inception module (DCCIM) shown in Fig. 5 to enhance the contextual inception features. Unlike in the receptive field block module [51], we employed a densely connected structure to transmit global contextual inception information and a residual module [31] to preserve original information. Using the DCCIM, we obtained enhanced features F_i^{dccim} as follows:

$$F_i^{dccim} = F_{DCCIM}\left(F_i^{dense}\right), \tag{11}$$

where F_i^{dccim} is the DCCIM output, and F_{DCCIM} denotes the DCCIM operation.

As Fig. 5 shows, the DCCIM contained three densely cascaded dilated convolutional layers. For these layers, a 1×1-kernel convolution was first used to reduce the number of channels to 16. For convolutional layer j (=1, 2, 3), two cascaded convolutions with kernels of $1 \times (2j + 1)$ and $(2j+1) \times 1$ and dilation rate of one were used. Subsequently, another 3×3-kernel convolution layer with a dilation rate of $(2j + 1)$ was applied. Next, we concatenated the features generated from the densely connected structure and dense features F_i^{dense}, and a 3×3-kernel convolution was used to change the channel size to 64.

Thus, F_{DCCIM} was formulated as follows:

$$F_{DCCIM}(f) = Conv_3\left(Cat\left(M\left(Conv_1(f)\right), f\right)\right), \tag{12}$$

where f is the input features, $Conv_1$ changes the channel size to 16, $Conv_3$ changes the channel size of the features to 64, and M denotes three densely cascaded convolutional layers, where the output of the previous layer is superimposed as the input of the next layer.

Finally, we used a skip-connection structure to integrate the two DFFM inputs (i.e., $F\ rgb\ i$ and F_i^t) and enhanced features $F\ dccim\ i$. We used the DCCIM again to further enhance the

TABLE I
RESULTS ON THE MFNET DATASET. EACH VALUE IN BOLDFACE INDICATES THE BEST RESULT FOR THE CORRESPONDING COLUMN

Methods	Color Cone		Person		Car Stop		Curve		Bike		Guardrail		Car		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
FRRN (3c)	34.7	32.5	53.0	46.1	21.6	19.1	34.0	27.1	65.1	53.0	0.0	0.0	80.0	71.2	36.2	30.5	47.1	41.8
FRRN (4c)	37.2	34.0	66.2	60.8	12.5	11.5	41.2	35.0	62.8	50.3	0.0	0.0	81.9	74.7	35.2	34.6	48.5	44.2
DFN (3c)	44.2	31.0	67.7	52.8	35.1	23.8	49.2	34.9	71.5	57.5	4.1	0.9	90.7	81.4	54.6	47.5	57.3	47.5
DFN (4c)	48.3	42.5	73.2	65.0	38.9	25.7	54.0	40.4	75.5	60.9	10.2	4.0	90.0	84.4	55.8	47.4	60.5	52.0
BiSeNet (3c)	49.6	43.3	65.0	54.3	32.3	26.2	32.1	25.7	75.0	61.4	3.2	0.9	90.0	84.5	48.1	40.5	54.9	48.2
BiSeNet (4c)	47.4	42.2	72.0	63.2	34.2	25.3	45.1	36.7	74.1	60.1	18.2	5.0	89.7	84.1	39.8	35.9	57.7	50.0
SHRNet (3c)	54.6	47.2	73.1	59.8	38.3	28.7	47.0	33.2	74.9	61.3	7.3	1.4	92.2	86.6	61.5	46.2	60.9	51.3
SHRNet (4c)	56.5	**49.8**	79.3	71.0	25.7	19.1	59.8	42.5	78.3	63.4	18.8	2.7	92.8	87.6	63.5	44.5	63.7	53.2
D-CNN	32.9	30.1	61.7	53.4	41.3	29.3	40.2	30.9	76.0	56.5	22.8	8.5	85.2	77.0	36.5	32.3	55.1	46.1
FuseNet	31.1	21.4	75.2	66.3	17.4	15.0	51.0	37.8	64.5	51.9	0.0	0.0	81.0	75.6	51.9	45.0	52.4	45.6
MFNet	30.3	25.2	67.0	58.9	12.5	9.9	36.2	29.9	53.9	42.9	0.1	0.0	77.2	65.9	30.0	27.7	45.1	39.7
RTFNet	45.5	29.1	79.3	70.3	38.5	29.8	60.7	45.3	76.8	62.7	0.0	0.0	93.0	87.4	74.7	55.7	63.1	53.2
FusSeg	**55.8**	46.9	81.4	71.7	29.1	22.7	**68.4**	**44.8**	**78.5**	**64.6**	63.7	6.4	93.1	**87.9**	66.4	**47.9**	70.6	54.5
Ours	54.7	48.7	**83.0**	**73.1**	**55.0**	**42.3**	59.7	44.0	76.9	61.7	**71.2**	**14.5**	**94.1**	86.5	**73.1**	47.4	**74.1**	**57.3**

features. Thus, we obtained DFFM output F_i^{fused} as follows:

$$F_i^{fused} = F_{DCCIM}\left(F_i^{dccim} \oplus Conv_1\left(F_i^{rgb} \oplus F_i^{t}\right)\right), \quad i = 3, 4, 5, \quad (13)$$

where $Conv_1$ changes the channel size of the features to 64.

F. Loss Function

The difficulty to classify a pixel is closely related to its position. Over a cluttered background, pixels near the edge are more inclined to incorrect classifications, whereas central pixels have a higher segmentation accuracy owing to the internal consistency of the semantic target. Instead of treating pixels equally, we handled them according to their characteristics. Specifically, we decoupled the original label into a binary and boundary label. For the binary label, we assigned the values of one and zero to foreground and background classes, respectively, to distinguish them. For the boundary label, a sliding window was used on the semantic label to determine the class of a region. If the semantic class within the sliding window was different, the center of the window was considered a boundary. Here, we set the size of the sliding window to 5×5.

We obtained semantic, binary, and boundary outputs, each with its corresponding loss. Therefore, training loss L was defined as the combination of the three loss functions:

$$L = \alpha_1 \ell_{semantic} + \alpha_2 \ell_{binary} + \alpha_3 \ell_{bound}, \quad (14)$$

where $\ell_{semantic}$, ℓ_{binary}, and ℓ_{bound} represent the semantic, binary, and boundary loss functions, respectively, and α_1, α_2, and α_3 are the corresponding loss weights that control the tradeoff between the loss functions and sum to one. According to experiments with different settings, we optimally set the weights of $\{\alpha_1, \alpha_2, \alpha_3\}$ as $\{0.4, 0.3, 0.3\}$ to refine the proposed GMNet (the choice of these weights is driven by the goal of having good results on all evaluation measures and by taking into account the numerical range that the single measures have at convergence). We directly used the binary cross-entropy to compute both ℓ_{binary} and ℓ_{bound}. The binary cross-entropy is a widely used loss function for binary segmentation and classification. It is defined as follows:

$$\ell_{CE}(S, G) = G \cdot \log S + (1 - G) \cdot \log(1 - G), \quad (15)$$

where S is the predicted map generated in the last two stages (i.e., S_2', S_3'), and G denotes the ground-truth label. However, the binary cross-entropy computes the loss for each pixel independently, neglecting the global image structure. To consider global information, we used the Lovász-softmax loss, as suggested in [52], to calculate $\ell_{semantic}$. This loss function enables a plug-and-play loss layer to directly optimize the mean intersection-over-union (mIoU) loss in neural networks

TABLE II
RESULTS FROM NIGHTTIME AND DAYTIME IMAGES. VALUES IN BOLDFACE INDICATE THE BEST RESULTS

Methods	Daytime		Nighttime	
	mAcc	mIoU	mAcc	mIoU
FRRN (3c)	45.1	40.0	41.6	37.3
FRRN (4c)	42.4	38.0	46.2	42.3
DFN (3c)	53.7	42.2	52.4	44.6
DFN (4c)	53.4	43.9	57.4	51.8
BiSeNet (3c)	52.1	44.5	50.3	45.0
BiSeNet (4c)	52.9	44.8	53.1	47.7
SegHRNet (3c)	59.7	47.2	55.7	49.1
SegHRNet (4c)	50.0	41.4	50.2	44.9
D-CNN	50.6	42.4	50.7	43.2
FuseNet	49.5	41.0	48.9	43.9
MFNet	42.6	36.1	41.4	36.8
RTFNet	60.0	45.8	60.7	54.8
FusSeg	62.1	47.8	67.3	54.6
Ours	**71.0**	**49.0**	**71.3**	**57.7**

TABLE III
RESULTS ON THE PST900 DATASET

Methods	Background		Hand-Drill		Backpack		Fire-Extinguisher		Survivor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Efficient FCN (3c)	99.81	98.63	32.08	30.12	60.06	58.15	78.87	39.96	32.76	28.00	60.72	50.98
Efficient FCN (4c)	99.80	98.85	48.75	38.58	69.90	67.59	76.45	46.28	38.86	35.06	66.75	57.27
CCNet (3c)	**99.86**	99.05	51.77	32.27	68.30	66.42	67.79	51.84	60.84	57.50	69.71	61.42
CCNet (4c)	99.59	98.74	54.09	51.01	75.96	72.95	88.06	73.80	49.45	33.52	73.43	66.00
ACNet	99.83	99.25	53.59	51.46	85.56	83.19	84.88	59.95	69.10	65.19	78.67	71.81
SA-Gate	99.74	99.25	89.88	81.01	**89.03**	79.77	80.70	72.97	64.19	62.22	84.71	79.05
RTFNet	99.78	99.02	7.79	7.07	79.96	74.17	62.39	51.93	78.51	70.11	65.69	60.46
PSTNet	-	98.85	-	53.60	-	69.20	-	70.12	-	50.03	-	68.36
Ours	99.81	**99.44**	**90.29**	**85.17**	89.01	**83.82**	**88.28**	**73.79**	**80.65**	**78.36**	**89.61**	**84.12**

and appropriately reflects global contextual information. More details about this loss function can be found in [52].

IV. RESULTS AND DISCUSSION

A. Datasets

In this study, two publicly available RGB-T datasets were employed to train and evaluate the proposed GMNet. One dataset was released by Ha *et al.* [14], consisting of urban scenes with common objects, namely bike, person, car, curve (road lanes), guardrail, car stop, bump, and color cone. An InfReC R500 camera was utilized to obtain 1,569 registered RGB-T images at a resolution of 480 × 640 pixels. From these, 749 nighttime images and 820 daytime images were obtained. The dataset contained nine manually labeled classes, including one background class (no label) and the eight classes of common objects. The other dataset was PST900 [48], which included 894 matched RGB-T images with pixel-level manual annotations of four classes from the DARPA Subterranean Challenge.

B. Training

A computer equipped with an Intel Core i5 processor and an NVIDIA TITAN RTX graphics card with 24 GB memory was used to train the proposed GMNet. Furthermore, the splitting strategy was the same as in [14]. Specifically, we selected 50% of the nighttime RGB-T images and 50% of the daytime RGB-T images to construct the training set. The remaining 25% of the nighttime RGB-T images and 25% of the daytime RGB-T images constituted the validation and test sets. We also applied flipping to augment the training set. The proposed GMNet was executed using PyTorch, and we initialized the GMNet encoders with the pretrained weights provided in PyTorch. We set the batch size to four for fitting the GPU memory and optimized the network by applying stochastic gradient descent. Additionally, the initial learning rate was set to 0.0005 and decayed exponentially. The training did not terminate until the loss stopped decreasing. The convergence of our training loss was depicted in Fig. 6.

C. Evaluation Measures

We used two evaluation measures to quantitatively estimate the semantic segmentation performance, namely accuracy per class, also known as recall, and intersection over union per class. The mean values across all the classes in a dataset for the two evaluation measures are denoted as *mIoU* and *mAcc*, respectively, and they are given by

$$mAcc_i = \frac{\sum_{k=1}^{K} P_{ii}^k}{\sum_{k=1}^{K} P_{ii}^k + \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{N} P_{ij}^k}, \tag{16}$$

$$mIoU_i = \frac{\sum_{k=1}^{K} P_{ii}^k}{\sum_{k=1}^{K} P_{ii}^k + \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{N} P_{ji}^k + \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{N} P_{ij}^k}, \tag{17}$$

where K and N ($N = 9$ here) are the numbers of test images and classes, respectively. For frame k, P_{ii}^k is the number of pixels for class i correctly classified as class i, P_{ji}^k is the number of pixels for class j incorrectly classified as class i, and P_{ij}^k is the number of pixels for class i incorrectly classified as class j.

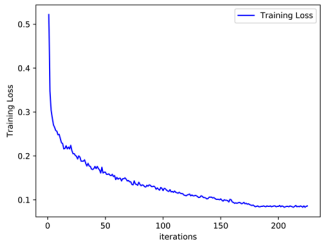

Fig. 6. Convergence of training loss in proposed GMNet.

D. Evaluation on MFNet Dataset

1) Overall Results: We compared the proposed GMNet with FuseSeg [16], RTFNet [15], MFNet [14], DFN [28], BiSeNet [29], SegHRNet [31], FuseNet [37], D-CNN [35], and FRRN [53] on the MFNet dataset. The results of MFNet, FuseNet, RTFNet, and FuseSeg were obtained from [16] to facilitate comparison. Table I lists the quantitative comparison results. The proposed GMNet with its ResNet-50 backbone outperformed the other methods with regard to both *mAcc* and *mIoU*.

2) Daytime and Nighttime Results: To further evaluate the methods, we tested them on daytime and nighttime RGB-T

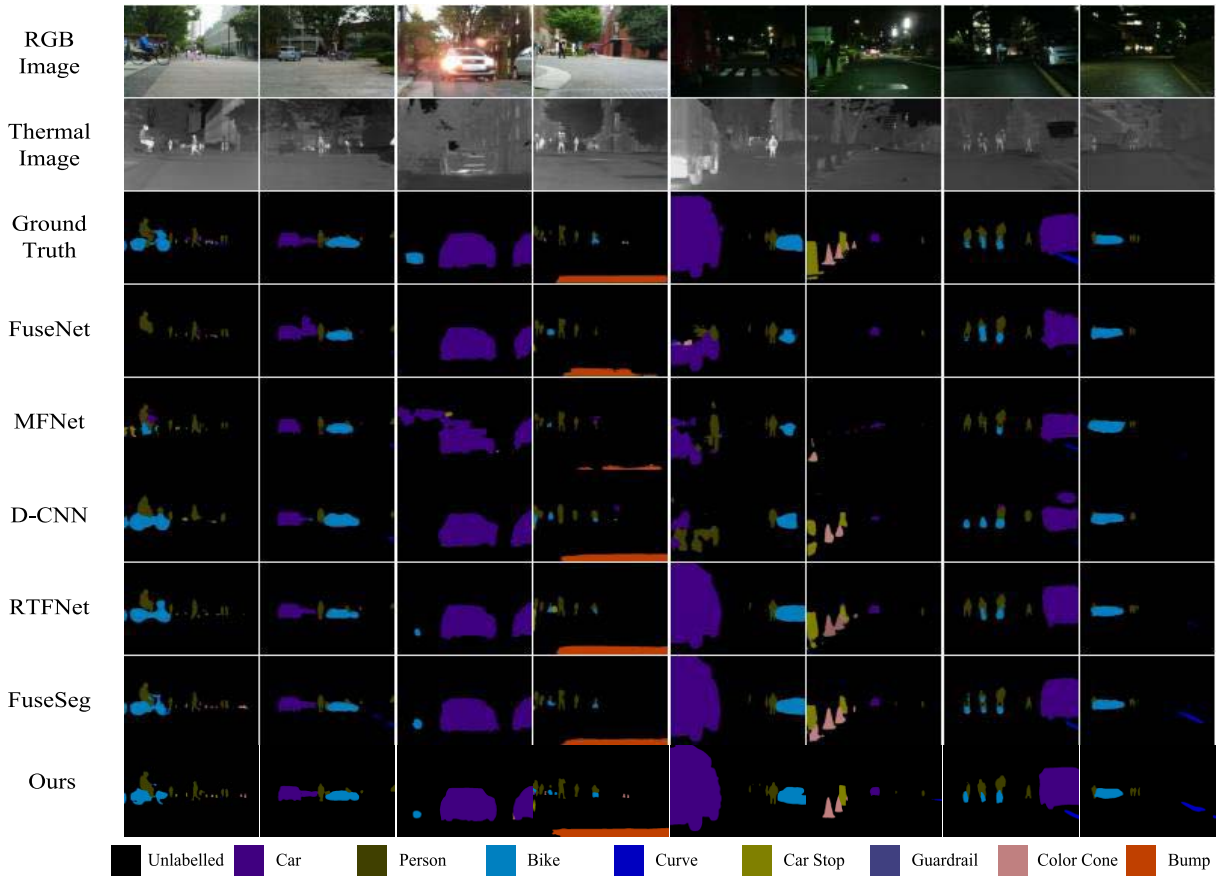

Fig. 7. Segmentation results of fusion modules in typical nighttime and daytime RGB-T images shown in the right four and left four columns, respectively. The proposed GMNet provides better segmentation under varying lighting conditions than the comparison networks.

TABLE IV

QUANTITATIVE RESULTS (%) ON THE TEST SET OF SUN RGB-D DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLDFACE

Methods	mAcc	mIoU
FCN	38.4	27.4
DeconvNet	32.3	22.6
D-CNN	36.9	27.2
FuseNet	38.0	29.4
SegNet	44.8	31.8
ACNet	-	48.1
RDFNet	**60.1**	47.7
FuseSeg	38.3	28.8
Ours	58.6	**48.7**

TABLE V

RESULTS OF ABLATION EXPERIMENTS FOR THE BACKBONES. THE PROPOSED GMNET USES THE RESNET-50 BACKBONE. VALUES IN BOLDFACE INDICATE THE BEST RESULTS

Variants	mAcc	mIoU
MFNet [14]	45.1	39.7
RTFNet [15]	63.1	53.2
FuseSeg [16]	70.6	54.5
VGG16	69.8	54.9
MobileV2	64.8	50.5
DenseNet161	**75.5**	53.1
ResNet18	72.7	55.9
ResNet34	71.2	54.4
ResNet101	74.2	56.4
ResNet152	74.3	56.6
ResNet50(Ours)	74.1	**57.3**

images. Table II and Fig. 7 show the comparative and visual segmentation results. The proposed GMNet outperformed most of the other methods. On daytime RGB-T images, possible registration errors [14] may misguide segmentation, likely explaining that some single-modal methods using RGB data outperformed the methods using four-channel RGB-T data. In practice, both RGB and thermal images carry beneficial information on daytime images, but spatial or temporal misalignments between the two modalities may provide misleading information that undermines the performance of semantic segmentation. On nighttime images, almost all the multimodal methods provided superior results when using four-channel information. This was reasonable as RGB data are less informative under poor lighting conditions (e.g., dimness and darkness), whereas thermal images are not susceptible to such conditions.

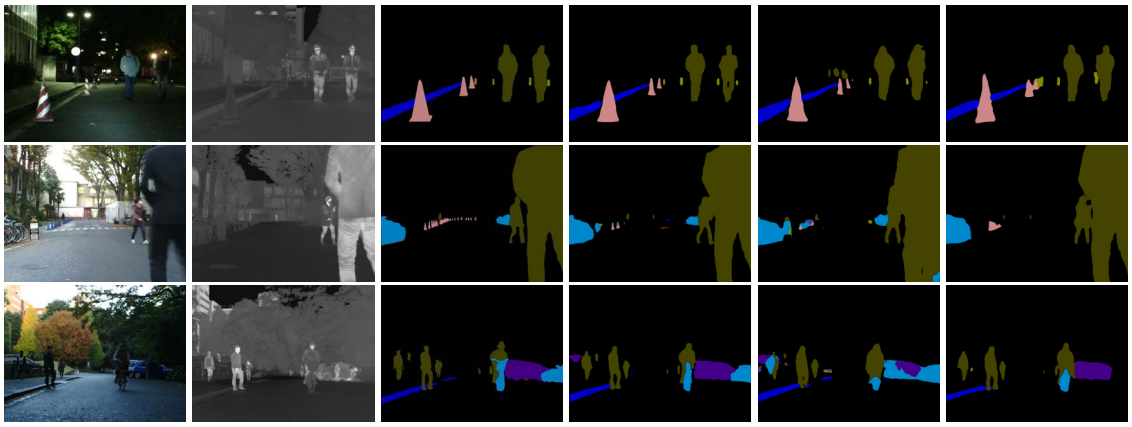

Fig. 8. Segmentation results of NTE (no thermal encoder network), NRE (no RGB encoder network), and complete GMNet.

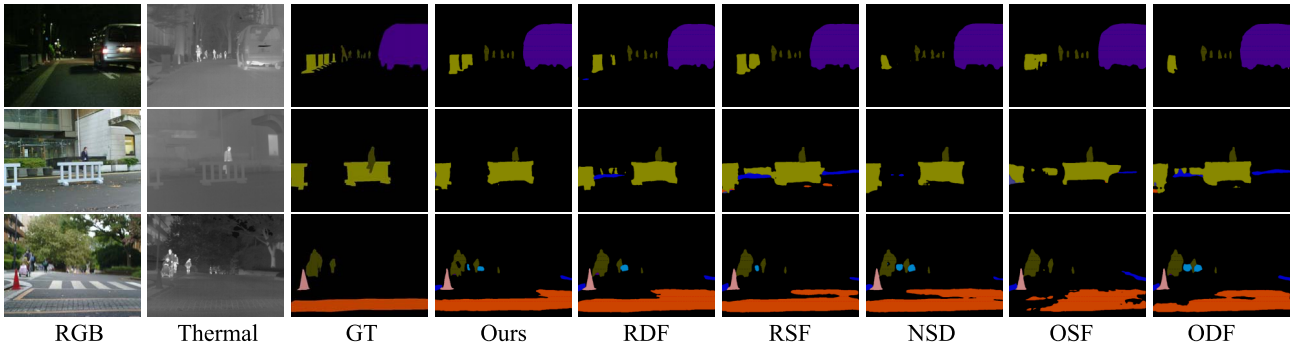

Fig. 9. Segmentation results of RDF (no DFFM), RSF (no SFFM), NSD (neither SFFM nor DFFM), OSF (SFFM only network), ODF (DFFM only network) variants, and complete GMNet.

TABLE VI

RESULTS OF ABLATION EXPERIMENTS FOR THE ENCODERS. VALUES IN BOLDFACE INDICATE THE BEST RESULTS

Variants	Ours	NTE	NRE
mAcc	**74.1**	68.2	67.1
mIoU	**57.3**	50.7	48.7

TABLE VII

RESULTS OF ABLATION EXPERIMENTS FOR FUSION STRATEGY. VALUES IN BOLDFACE INDICATE THE BEST RESULTS

Variants	Ours	RDF	RSF	NSD	OSF	ODF
mAcc	74.1	70.5	**75.5**	73.4	69.6	72.0
mIoU	**57.3**	57.2	54.0	53.4	54.1	55.8

E. Evaluation on PST900 Dataset

To evaluate the effectiveness of the proposed GMNet, we also tested it on the PST900 RGB-T dataset [48]; Table III lists the evaluation results (EFFicient FCN [33], CCNet [34], ACNet [12], SA-Gate [54], RTFNet [15], and PSTNet [48]). The proposed GMNet outperformed the SOTA MFNet, RTFNet, and PSTNet by a large margin, demonstrating the generalization ability of GMNet.

F. Generalization to RGB-D Data

To further verify the generalization ability of GMNet, we used the SUN RGB-D indoor scene analysis baseline dataset [55] to train and test the proposed GMNet. Table VI lists the comparison results (FCN [20], DeconvNet [22], FuseNet [37], SegNet [23], ACNet [12], RDFNet [39], and FuseSeg [16]). The proposed GMNet achieved a high performance indicating that it can be generalized to RGB-D data.

G. Ablation Studies

1) Scalability: To validate the scalability of the proposed GMNet, we compared the performance of several commonly used structures (e.g., VGG16 [21], MobileNetV2 [56], DenseNet161 [49], and different ResNet configurations) as backbones, obtaining the results listed in Table V. In general, the proposed GMNet outperformed the SOTA MFNet, RTFNet, and FuseSeg with all the ResNet backbones and achieved a performance comparable to the other three backbones (i.e., VGG16, MobileNetV2, and DenseNet161), indicating the strong scalability of GMNet. In particular, ResNet-50 achieved the highest *mIoU* among the evaluated backbones, and considering its accuracy and speed, ResNet-50 seemed to be the best backbone for GMNet.

2) Multimodal Information: We deleted the thermal encoder of GMNet to evaluate its performance without either the thermal (NTE) or RGB (NRE) information. In these variants, fusion was not performed because only one encoder

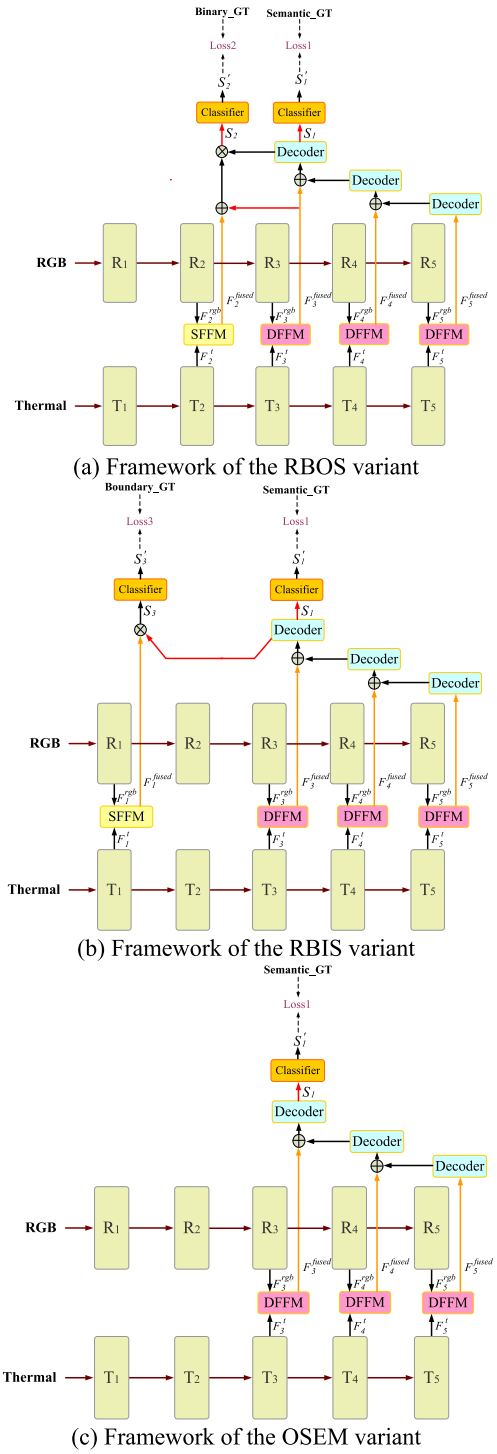

Fig. 10. Ablation study of supervised learning.

remained in GMNet. Table VI lists the evaluation results, and Fig. 8 shows the segmentation results. The complete GMNet outperformed both NTE and NRE, demonstrating the high performance of the multimodal structure. The superior performance was expected because only RGB information was misleading under adverse lighting conditions; whereas, only thermal information lacked detailed and textured information that could be helpful for segmentation.

3) Feature Fusion: For the ablation experiment of the fusion strategy, we compared the fusion strategies and fusion modules in five GMNet variants. The RDF variant replaced the DFFMs in the last three layers by elementwise summations. Similarly, the RSF variant replaced the SFFMs in the first two layers by elementwise summations. The NSD variant replaced the SFFMs and DFFMs by elementwise summations. Therefore, this variant had no fusion modules. The OSF variant replaced the DFFMs in the last three layers by SFFMs to perform shallow feature fusion only. Finally, the ODF variant replaced the SFFMs in the first two layers by DFFMs to perform deep feature fusion only.

Table VII and Fig. 9 show the evaluation and segmentation results of this ablation study. The complete GMNet with its different fusion modules exhibited the highest performance, confirming the effectiveness of the proposed fusion strategy. The results obtained from RDF, RSF, and NSD indicated that the SFFM and DFFM improved the segmentation performance. This performance was clearly superior to a simple summation between two encoders, which was not sufficient and failed to fully use the complementarity between modalities. In contrast, SFFM and DFFM suitably aggregated RGB and thermal information while enhancing feature fusion. The results obtained from OSF and ODF indicated that only SFFM or DFFM could not achieve the high performance of the combined fusion strategy, confirming our hypothesis that uniform fusion might not be suitable for multilevel networks. In fact, the SFFM enhanced spatial details owing to its attention structure, while the DFFM suitably aggregated and improved contextual semantic information owing to its multiscale perception ability. Thus, it was reasonable to include the SFFM in shallow layers and the DFFM in deep layers.

4) Supervised Learning: For the ablation study of supervised learning, we compared the GMNet supervision strategy with three variants. The RBOS variant removed the boundary supervision in the third decoding stage while preserving binary supervision and semantic supervision (see Fig. 10[a]). The RBIS variant removed the binary supervision in the second decoding stage while preserving boundary supervision and semantic supervision (see Fig. 10[b]). Finally, the OSEM variant removed both the boundary supervision and binary supervision, preserving only the semantic supervision in the first decoding stage; furthermore, the OSEM variant also removed SFFM (see Fig. 10[c]).

Table VIII and Fig. 11 show the evaluation and segmentation results of this ablation study. Table VIII shows that the *mIoU* of the complete GMNet exceeded that of the variants by a large margin, and *mAcc* provided satisfactory results. In the first row of Fig. 11, the segmentation performance was not satisfactory for some small and distant objects (e.g., person and bike) when removing boundary supervision (RBOS

TABLE VIII

RESULTS OF ABLATION EXPERIMENTS FOR SUPERVISION. VALUES IN BOLDFACE INDICATE THE BEST RESULTS

Variants	Ours	RBOS	RBIS	OSEM
mAcc	74.1	**74.4**	72.1	61.4
mIoU	**57.3**	53.9	53.8	50.5

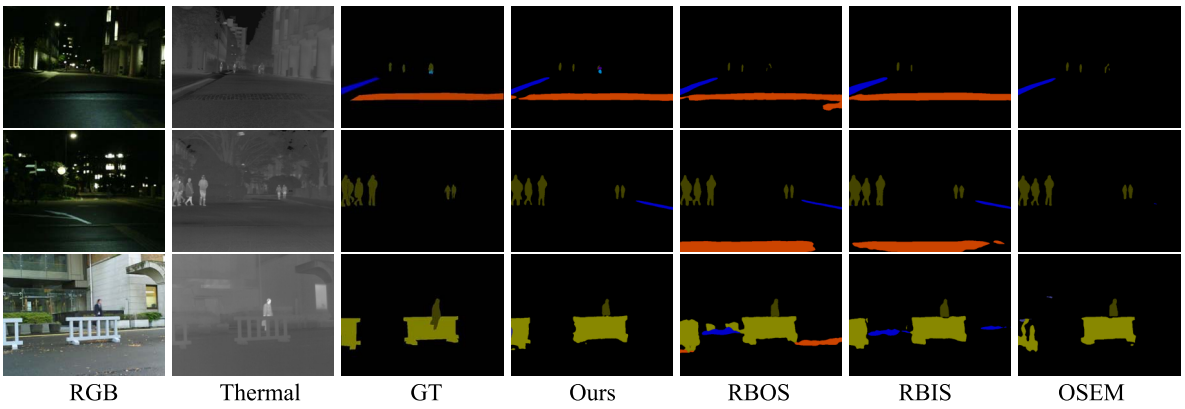

Fig. 11. Segmentation results of RBOS (no boundary supervision), RBIS (no binary supervision), OSEM (segmentation supervision only) variants, and complete GMNet.

variant), binary supervision (RBIS variant), or both (OSEM variant). In the second row of Fig. 11, incorrect semantic segmentation occurred on some regions, such as around the bump. Furthermore, the semantic segmentation of the bump failed completely when only semantic supervision (OSEM variant) was preserved.

V. Conclusion

In this study, we developed GMNet for RGB-T urban scene semantic segmentation. We used thermal data to complement RGB data, combined the cross-model graded features using two distinct fusion modules, and used three supervision strategies for optimization. Using GMNet, we aimed to achieve superior results under adverse lighting conditions (e.g., dimness and darkness), and our experimental results demonstrated the superiority of the proposed method over SOTA methods. Comprehensive ablation studies confirmed the effectiveness of the adopted data fusion strategy and multilabel supervision. Remarkably, the proposed GMNet was backbone-independent, making it promising and challenging for research on other topics, including super-resolution, object detection, and classification. Evaluations of the GMNet on the SUN RGB-D indoor scene dataset demonstrated its excellent generalization ability. As part of our future work, in addition to more applications, we will explore other loss functions to further improve the performance of GMNet.

References

[1] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 9080–9089.

[2] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[3] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Trans. Image Process.*, vol. 29, pp. 3520–3533, 2020.

[4] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[5] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for RGB thermal urban road scene parsing," *IEEE Trans. Multimedia*, early access, Jun. 7, 2021, doi: 10.1109/TMM.2021.3086618.

[6] J. Wu, W. Zhou, T. Luo, L. Yu, and J. Lei, "Multiscale multilevel context and multimodal fusion for RGB-D salient object detection," *Signal Process.*, vol. 178, Jan. 2021, Art. no. 107766.

[7] W. Zhou, S. Pan, J. Lei, and L. Yu, "TMFNet: Three-input multilevel fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, Aug. 12, 2021, doi: 10.1109/TETCI.2021.3097393.

[8] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Global and local-contrast guides content-aware fusion for RGB-D saliency prediction," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 6, pp. 3641–3649, Jun. 2021.

[9] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 3, 2021, doi: 10.1109/TCSVT.2021.3077058.

[10] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, early access, May 6, 2021, doi: 10.1109/TMM.2021.3077767.

[11] W. Zhou, W. Liu, J. Lei, T. Luo, and L. Yu, "Deep binocular fixation prediction using a hierarchical multimodal fusion network," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jan. 12, 2021, doi: 10.1109/TCDS.2021.3051010.

[12] 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)*, Taipei, Taiwan, Sep. 2019, pp. 1440–1444.

[13] W. Zhou, J. Wu, J. Lei, J.-N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder," *IEEE Trans. Multimedia*, early access, Sep. 20, 2020, doi: 10.1109/TMM.2020.3025166.

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

[15] 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.

[16] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021, doi: 10.1109/TASE.2020.2993143.

[17] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 3127–3135.

[18] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "IRFR-Net: Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 20, 2021, doi: 10.1109/TNNLS.2021.3105484.

[19] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 2021, doi: 10.1109/TGRS.2021.3109626.

[20] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[23] 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.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[25] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147
[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[27] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.
[28] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1857–1866.
[29] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.
[31] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*. [Online]. Available: http://arxiv.org/abs/1904.04514
[32] S. Kong and C. Fowlkes, "Pixel-wise attentional gating for scene parsing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1024–1033.
[33] J. Liu, J. He, J. Zhang, J.-S. Ren, and H. Li, "EfficientFCN: Holistically-guided decoding for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.
[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[35] S. Huo, Y. Zhou, and S.-Y. Kung, "Salient object detection via a linear feedback control system," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4257–4261.
[36] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 7253–7262, doi: 10.1109/ICCV.2019.00735.
[37] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Computer Vision—ACCV*. Cham, Switzerland: Springer, 2017, pp. 213–228.
[38] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Comput. Vis. (ECCV)*, Sep. 2018, pp. 135–150.
[39] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.
[40] X. Gao, J. Yu, and J. Li, "RGBD semantic segmentation based on global convolutional network," in *Proc. 4th Int. Conf. Robot., Control Automat.*, Jul. 2019, pp. 192–197.
[41] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*. [Online]. Available: http://arxiv.org/abs/1806.01054
[42] W. Zhou, J. Yuan, J. Lei, and T. Luo, "TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation," *IEEE Intell. Syst.*, early access, Jun. 10, 2020, doi: 10.1109/MIS.2020.2999462.
[43] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, "RFBNet: Deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation," 2019, *arXiv:1907.00135*. [Online]. Available: http://arxiv.org/abs/1907.00135
[44] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Comput. Vis. (ECCV)*, 2020, pp. 561–577.
[45] D. Lin, R. Zhang, Y. Ji, P. Li, and H. Huang, "SCN: Switchable context network for semantic segmentation of RGB-D images," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1120–1131, Mar. 2020.
[46] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational context-deformable ConvNets for indoor scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3991–4001.
[47] G. Zhang, J.-H. Xue, P. Xie, S. Yang, and G. Wang, "Non-local aggregation for RGB-D semantic segmentation," *IEEE Signal Process. Lett.*, vol. 28, pp. 658–662, 2021.
[48] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Paris, France, May 2020, pp. 9441–9447.
[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] Y. Zhu, J. Mu, H. Pu, and B. Shu, "FRFB: Integrate receptive field block into feature fusion net for single shot multibox detector," in *Proc. 14th Int. Conf. Semantics, Knowl. Grids (SKG)*, Guangzhou, China, Sep. 2018, pp. 173–180.
[52] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4413–4421.
[53] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3309–3318.
[54] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 561–577.
[55] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 564–571.
[56] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.

Wujie Zhou (Member, IEEE) is currently an Associate Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China. He is also a Postdoctoral Fellow with the Institute of Information and Communication Engineering, Zhejiang University, Zhejiang. His research interests include multimedia signal processing and communication. He is a Reviewer for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON BROADCASTING, the IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, *Information Sciences*, *Neurocomputing*, and *SPIC*.

Jinfu Liu received the B.S. degree from the College of Electronic Information Engineering, Tianjin University, Tianjin, China, in 2015. He is currently pursuing the M.S. degree with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China. His research interests include multimedia signal processing and communication.

Jingsheng Lei received the B.S. degree in mathematics from Shanxi Normal University in 1987 and the M.S. and Ph.D. degrees in computer science from Xinjiang University in 2000 and 2003, respectively. He is currently a Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China. His research interests include machine learning, data mining, pattern recognition, and cloud computing. He is a member of the Artificial Intelligence and Pattern Recognition Technical Committee of China Computer Federation and the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence.

Lu Yu (Member, IEEE) received the B.Eng. degree in radio engineering and the Ph.D. degree in communication and electronic systems from Zhejiang University, Hangzhou, China, in 1991 and 1996, respectively. She is currently a Professor with the Institute of Information and Communication Engineering, Zhejiang University. Her current research interests include video coding, multimedia communication, and relative application-specific integrated circuit design.

Jenq-Neng Hwang (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California.

In 1989, he joined the Department of Electrical and Computer Engineering (ECE), University of Washington, Seattle, WA, USA, where he has been promoted to a Full Professor in 1999. He was the Associate Chair for research from 2003 to 2005 and from 2011 to 2015. He is currently the Associate Chair for global affairs and international development with the ECE Department. He is the Founder and the Co-Director of the Information Processing Laboratory, which received several AI City Challenges awards. He has written more than 330 journals, conference papers, and book chapters in the areas of machine learning, multimedia signal processing, and multimedia system integration and networking. He has authored a textbook *Multimedia Networking: From Theory to Practice* (Cambridge University Press). He has close working relationship with the industry on multimedia signal processing and multimedia networking. He is a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He is also a member of the Multimedia Technical Committee of the IEEE Communication Society and the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He received the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He was the Program Co-Chair of the ICASSP 1998 and the ISCAS 2009. He has served as the Program Co-Chair for the IEEE ICME 2016. He was the Society's Representative of the IEEE Neural Network Council from 1996 to 2000. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON IMAGE PROCESSING, and the *IEEE Signal Processing Magazine*. He is on the Editorial Board of the *ZTE Communications*, *ETRI*, *IJDMB*, and *JSPS* journals.