NLFNet: Non-Local Fusion Towards Generalized Multimodal Semantic Segmentation across RGB-Depth, Polarization, and Thermal Images

Ran Yan[1], Kailun Yang[2], and Kaiwei Wang[3]

***Abstract*— In recent years, intelligent driving navigation has made considerable progress, and semantic segmentation is one of the most advanced scene perception methods. At present, traditional semantic segmentation methods can use RGB images for detection of obstacles that are clearly visible in outdoor scenes. However, in the face of complex realistic driving scenes, RGB images cannot provide sufficient information. We need some other modal information to supplement the RGB information. In this paper, we propose *Non-Local Fusion Network (NLFNet)*, which is a semantic segmentation network that can selectively fuse multimodal input information in an adaptive manner. It can use complementary information collected by different optical sensors to extract effective features for fusion. Thereby, it improves the segmentation accuracy of the network and solves the problem of object recognition in various challenging real-world scenes. We conduct comprehensive experiments to verify the effectiveness and generalization ability of the framework across *RGB-Depth*, *RGB-Polarization*, and *RGB-Thermal* image semantic segmentation, which is especially suitable for autonomous driving and robot vision applications.**

I. Introduction

With the development of deep learning and Convolutional Neural Networks (CNNs) [1], the perception and understanding of outdoor scenes have become a hot topic for intelligent driving cars and intelligent mobile robots for navigation assistance. Image semantic segmentation is the basic task of computer vision. It aims to assign a semantic category to each pixel in the image, that is, to identify and classify objects at the pixel level. At present, many neural networks for semantic segmentation have been proposed, such as FCN [2], PSPNet [3], U-Net [4], *etc.* Thanks to the architectural advances and the emergence of large-scale datasets, these networks can accurately segment outdoor scenes under favorable environmental and lighting conditions, *e.g.*, they have a good distinction between *cars* and *backgrounds* in general road-driving street scenes (see Fig. 1).

The above-mentioned networks all use the information of RGB images for semantic segmentation, because RGB images can provide color information and texture information contained in the target object, and help the neural network of deep learning to recognize and segment the input image. But this also has certain flaws. When the object and the background have similar colors and textures, it is difficult for the neural network to distinguish them completely. For example, in detecting obstacles with high reflectivity such as *glass* and *metal*, the existing semantic segmentation methods are limited because these objects are difficult to detect, as shown in Fig. 1, and the existence of these obstacles is more dangerous for navigation assistance. They may pose potential dangers to self-driving cars and intelligent robots. In addition, under adverse weather conditions, the use of traditional detection methods will also reduce the accuracy of object segmentation. This has aroused increasing attention and research interests in the field of computer vision and intelligent robots [5], [6].

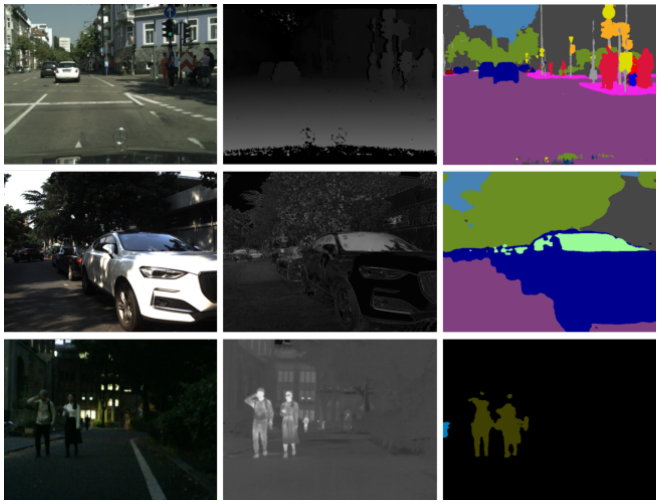

Fig. 1. Examples of images and recognition results: (a) RGB image, (b) Multimodal image (From top to bottom: depth image from Cityscapes [7], polarization image from ZJU-RGB-P [8], and thermal image from MFNet [9]), (c) Segmentation result using only RGB image.

When the color and texture information of the RGB image are not enough for the network to use for semantic segmentation, it is desired to combine the rich information of other modalities to supplement, so as to perform more accurate semantic segmentation of the target image. Considering the modal sensors that contain useful optical information, currently there are mainly color sensors, depth sensors, infrared sensors, polarization sensors, and event cameras, which contain various types of optical information [10]. For example, the depth image contains more object position and contour information, the polarization image can obtain the special polarization degree and polarization angle of the

This research was granted from ZJU-Sunny Photonics Innovation Center (No. 2020-03). This research was also funded in part through the AccessibleMaps project by the Federal Ministry of Labor and Social Affairs (BMAS) under the Grant No. 01KM151112 and and in part by the University of Excellence through the "KIT Future Fields" project. This research was also supported in part by Hangzhou SurImage Technology Company Ltd.

[1] R. Yan is with State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, China

[2] K. Yang is with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany kailun.yang@kit.edu

[3] K. Wang is with National Optical Instrumentation Engineering Technology Research Center, Zhejiang University, China wangkaiwei@zju.edu.cn

objects different from the background, and the infrared image can provide the thermal information of the objects and so on.

In this paper, we propose *Non-Local Fusion Network (NLFNet)*, a semantic segmentation network that can adaptively fuse multimodal input data, and use the multimodal image information collected by different optical sensors for effective feature extraction and fusion. We conduct extensive experiments, respectively fusing RGB information and depth information on the Cityscapes dataset [7], fusing RGB information and polarization information on the ZJU-RGB-P dataset [8], and fusing RGB information and thermal infrared information on the dataset provided in the work of MFNet [9]. The results show that NLFNet can effectively integrate and fuse multimodal information, significantly improving the accuracy of semantic segmentation, and it has a strong generalization capacity. In summary, this work delivers the following contributions:

- We propose NLFNet, which is a semantic segmentation network that effectively integrates multimodal image data. Compared with a single RGB network architecture, the segmentation of various objects are enhanced.
- The proposed network adaptively extracts complementary features of different modal input images, uses dependency information with long-range context priors, and improves accuracy of semantic segmentation.
- We conduct extensive experiments on different multimodal datasets, and comprehensively analyze the effectiveness and generalization ability of NLFNet in a wide variety of outdoor scenes.

II. Related Work

A. Semantic Segmentation Network

Since Fully Convolution Network (FCN) [2] implemented end-to-end pixel-level object classification, Convolutional Neural Networks (CNN) have developed rapidly in the field of semantic segmentation. SegNet [11] proposed an encoder-decoder network structure based on VGG16 [12]. U-Net [4] adds skip connections between the encoder and the decoder to merge low-level and high-level feature information. Deep network structures such as PSPNet [3] and DeepLab [13] construct a multi-scale representation, which increases the receptive field of the model. SENet [14] proposed a channel attention method and HANet [15] explored a height-driven context prior. DANet [16] and ECANet [17] devised variants of the non-local block [18] to capture long-range contextual dependency. Moreover, SETR [19], SegFormer [20], and Trans4Trans [21] revisited scene parsing from a sequence-to-sequence perspective, whereas MaskFormer [22] views semantic segmentation as a mask classification problem.

In addition, some networks have been improved towards real-time predictions like ERFNet [23] and SwiftNet [24]. They have applied early downsampling, filter decomposition, multi-branch structure, and ladder-style upsampling methods to effectively improve the efficiency of the network. Networks such as ACNet [25] and RFNet [26] use attention mechanisms and bridge connections, so that the semantic segmentation model can be executed quickly with multi-level feature aggregation. In this work, we leverage non-local mechanism and design an attention-based fusion architecture for multimodal semantic segmentation generalizable across various image modalities.

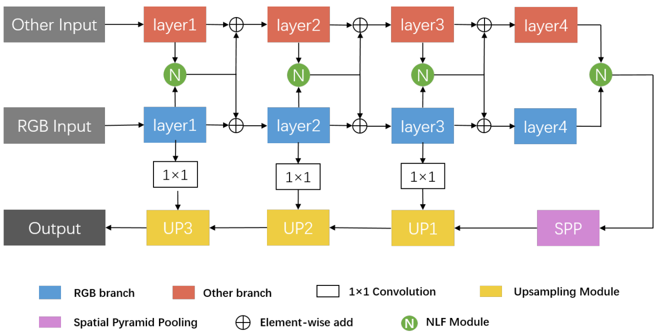

Fig. 2. Overview of NLFNet: the proposed network architecture based on non-local fusion for multimodal semantic segmentation.

B. Multimodal Fusion Semantic Segmentation

Although existing networks based only on RGB information have made some progress in the architecture, due to the limited input information, they will have certain restrictions on the segmentation performance of the image. In some complex environments or under challenging conditions, it is necessary to increase the richness of input information. Therefore, semantic segmentation network based on multimodal sensor data fusion has been widely researched. FuseNet [27] introduced the depth information of the environment on the basis of RGB images, and performed the fusion of RGB image and depth image features in the middle layer of the encoder and decoder. HeatNet [28] used infrared information as auxiliary information and adversarial training strategies to improve the performance of the model at night, which is challenging for RGB semantic segmentation methods [29], [30]. EAFNet [8] incorporated the polarization information of color polarization images to train the model. ISSAFE [6] leveraged event data and perform dense-to-sparse fusion to capture dynamic context information for improving semantic segmentation in extreme accident scenes. Moreover, there are many specialized RGB-D [31], [32] and RGB-T [33], [34], [35], [36] semantic segmentation methods. These networks improve the segmentation performance of objects in certain scenes, but it is desirable to design a robust fusion network that can effectively fuse different multimodal data.

III. Methods

In this section, we introduce the architecture of our designed Non-Local Fusion Network (NLFNet) for robust semantic segmentation across different modality combinations. In addition, we describe the characteristics of different multimodal datasets.

A. Network Architecture

Inspired by efficient networks such as SwiftNet [24] and RFNet [26], our NLFNet uses an encoder-decoder structure for semantic segmentation. The entire network architecture

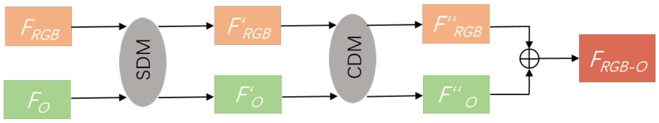

Fig. 3. Schematic diagram of Non-Local Fusion (NLF) Module.

of NLFNet is shown in Fig. 2. In the encoder part, we have two independent branches. We choose ResNet-18 [1] as the backbone of each network branch, because it has an appropriate depth and the residual structure, and at the same time has high computational efficiency. The two branches respectively perform downsampling and extract the latent features of the RGB images and the other modal images, and they are merged with the fusion operations. After obtaining the fused features, we use the Spatial Pyramid Pool (SPP) module [3], [26] to expand effective receptive fields and generate feature maps with more global contextual information, which are critical for accurate semantic segmentation. Finally, we perform the corresponding operations of the decoder to gradually restore and upsample the semantically-rich visual features from the coarse resolution to the input resolution. With reference to SwiftNet [24], we leverage three efficient upsampling modules, and merge the information of the RGB branch through skip connections. The 1×1 convolutions used in the upsampling process can connect the key elements of feature maps between deep and shallow layers to enhance detail sensitivity in the final semantic prediction and thereby improve the segmentation accuracy.

Inspired by Non-Local block [18] and NANet [32], we construct a Non-Local Fusion (NLF) module to integrate complementary information learned from the RGB branch and the other branch to achieve multi-level fusion of feature maps. The NLF module is mainly divided into two steps, including two sub-modules, as shown in Fig. 3. First, we establish long-range contextual dependency of the RGB branch and the other modal branch in space, by using a module termed Spatial Dependency Module (SDM). Then, we use a Channel Dependency Module (CDM) to establish cross-modal dependency on the two feature maps from different branches in the channel dimension.

For SDM, as shown in Fig. 4, for each spatial position in the original RGB image and the other modal image, it first performs global average pooling along the horizontal and vertical directions to extract the non-local information of the features, as depicted in the following equations:

$$F^{w}(c,1,j)=\frac{1}{H}\sum_{i=0}^{H-1}F(c,i,j) \tag{1}$$

$$F^{h}(c,i,1)=\frac{1}{W}\sum_{j=0}^{W-1}F(c,i,j) \tag{2}$$

After that, in the horizontal and vertical directions, we use 3×1 and 1×3 convolutions to expand the receptive fields respectively. Subsequently, the features are expanded to the original dimensions via upsampling, so we can obtain global width- and height-driven features F^{w}_{RGB-O} and F^{h}_{RGB-O}, as it can be seen in Fig. 4.

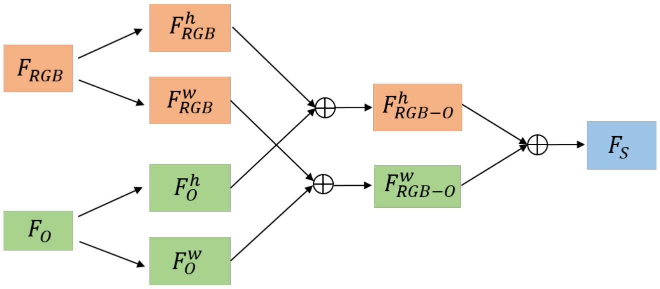

Fig. 4. Spatial Dependency Module (SDM).

By adaptively fusing non-local features including RGB features and O features (denoting the Other modality), a feature F_S that contains long-range dependencies in space is established. Finally, by fusing F_S with the original input features F_{RGB} and F_O via feature map addition, respectively, the integration of local RGB-O features and non-local RGB-O features is realized. Thereby, each spatial position of the original F_{RGB} feature can establish relationship with different positions of the F_O feature.

For CDM, as shown in Fig. 5, F'_{RGB} and F'_O outputs $\in\mathbb{R}^{H\times W\times C}$ from the SDM module are concatenated along the channel dimension to obtain the merged feature map $\in\mathbb{R}^{H\times W\times 2C}$, and then global average pooling is performed to obtain a squeezed feature map $\in\mathbb{R}^{1\times1\times 2C}$. It will be adaptively transformed into two independent embeddings $\in\mathbb{R}^{1\times1\times C}$ via fully conneted layers. Subsequently, the dependency weights W_{RGB} and W_O are obtained via a $Sigmoid$ activation layer. Finally, the output fused feature F_{RGB-O} is selectively obtained by associating the weights with the input features, as depicted in the following equation:

$$\begin{aligned}F_{RGB-O}&=F''_{RGB}\oplus F''_O\\&=\left(F'_{RGB}\otimes W_{RGB}\right)\oplus\left(F'_O\otimes W_O\right)\end{aligned} \tag{3}$$

In this way, we extract the nonlinear interactions between the cross-modal channels and establish non-local contextual dependencies between different modalities.

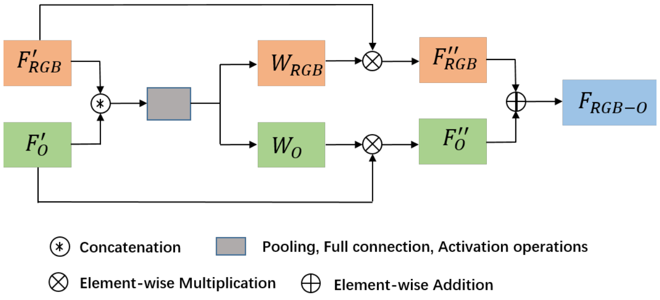

Fig. 5. Channel Dependency Module (CDM).

B. Multimodal Dataset

At present, the mainstream semantic segmentation methods mostly use RGB images as input information, because different types of objects have different color and texture information in RGB images, which can help the network to perform deep learning. Although the use of existing RGB images for training has a good semantic segmentation effect, the performance of image segmentation will be limited in

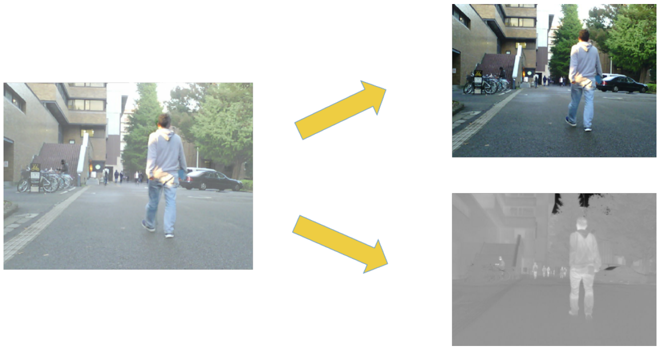

Fig. 6. Preprocessing of thermal image.

some complex scenes or under some challenging conditions. Therefore, it is necessary to use multimodal datasets and provide other forms of complementary data to enrich the input information of the semantic segmentation network.

RGB-Depth Fusion. Depth images can help the network to obtain the position and contour information of the objects, and the fusion of RGB information and depth information can better distinguish objects with different spatial positions. Cityscapes [7] is a common multimodal dataset, which contains outdoor street-view data of multiple cities. It provides 5000 sets of RGB-D images. By using the semi-global matching algorithm [37], disparity images can be obtained. It contains fine pixel-level annotations of 19 categories, with a resolution of 2048×1024. We use 2975 images from the training set for training and 500 images from the validation set for evaluation.

RGB-Thermal Fusion. At night or in places with insufficient light, objects and backgrounds in RGB images have similar color information and are difficult to distinguish. Thermal images can provide special infrared characteristics of objects, helping to segment objects such as *people* and *cars* at night. The RGB-Thermal dataset provided by the work of MFNet [9] contains 1569 images (820 images were taken during the day and 749 images were taken at night). 8 types of obstacles are marked with a resolution of 640×480. The training set includes 50% of the daytime images and 50% of the nighttime images, whereas the validation set contains 25% of the daytime images and 25% of the nighttime images [9]. Through preprocessing, the RGB image and thermal infrared image are extracted from the combined 4-channel image, as shown in Fig. 6.

RGB-Polarization Fusion. High-reflectivity objects such as glass and cars have RGB information that are easily confused with the environment. In a polarized image, objects with high reflectivity have a special polarization angle, which can be used to better segment these objects. The public dataset ZJU-RGB-P developed in our previous work [8] contains 394 groups of matched color images, each group has 4 images with different polarization angles. Including 9 categories of objects, the resolution is 1224×1024. We use 344 images from the training set for training, and the other 50 images as the validation set. As shown in Fig. 7, through preprocessing, the polarization degree image AoLP can be obtained via the following transformation:

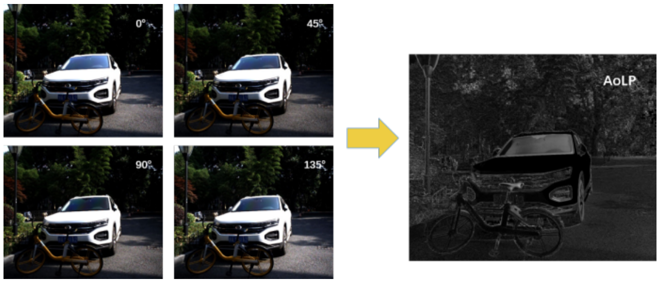

Fig. 7. Preprocessing of Polarization image: (a) RGB images at four directions and (b) Polarization image.

$$AoLP = \frac{1}{2} arctan\Big(\frac{I_0 - I_{90}}{I_{45} - I_{135}}\Big), \tag{4}$$

where I denotes the intensity image with the corresponding polarization angle.

IV. Experiments

In this section, we conduct a comprehensive analysis through extensive experiments.

A. Implementation Details

In this work, we utilize the three multimodal semantic segmentation datasets mentioned in the previous section. We perform data augmentation operations on all three datasets. Specifically, we first rescale the image with a random factor between 0.75 and 1.5, then randomly crop the images with a crop size of 480×480 , and finally perform a random horizontal flip.

We use CUDA 10.0, CUDNN 7.6.0, PyTorch 1.1 and an NVIDIA GeForce GTX 1080Ti GPU for model training. We use the pre-training weights on ImageNet [38] to initialize the ResNet-18 in the RGB branch and the O branch, and use the Adam optimizer [39] to optimize the learning rate. We set the initial learning rate to 4×10^{-4}. The cosine annealing learning rate scheduling strategy is used to adjust the learning rate, and the minimum value in the last epoch is 1×10^{-6}. In order to prevent overfitting, we use L2 weight regularization, and set the weight decay to 1×10^{-4}. We trained all models for 200 epochs with a batch size of 8. We use mean Intersection of Union (mIoU) to evaluate the segmentation accuracy of models, which represents the ratio of the intersection and union of the inference result and the ground truth.

B. Results and Analysis

Ablation Study. As shown in Table I, we conduct ablation studies on the ZJU-RGB-P dataset [8] with different networks to explore the effect of changes in the network architecture and fusion schemes on the segmentation accuracy. The single RGB method means that a single-branch network model SwiftNet is used and only RGB images are used as input information, achieving an mIoU of 80.3%. The single P method means that only the polarized image information is used. Compared with the single RGB method, mIoU is reduced to 73.5%. This is because the RGB image contains more effective information than the polarized image. In the

TABLE I

PERFORMANCE OF NLFNET ON THE ZJU-RGB-P VALIDATION DATASET [8] WITH DIFFERENT DESIGN CHOICES.

Network	Polarization	Dual-branch	RGB-P fusion	Element-wise add	mIoU(%)
SwiftNet (Single RGB)	✗	✗	✗	✗	80.3
SwiftNet (Single P)	✗	✗	✓	✗	73.5
SwiftNet (RGB-P-Stack)	✓	✗	✓	✗	80.2
NLFNet	✓	✓	✗	✓	84.4

TABLE II

ACCURACY ANALYSIS ON ZJU-RGB-P DATASET [8] INCLUDING PER-CLASS ACCURACY IN IOU (%) AND MEAN IOU (MIOU).

Network	Building	Glass	Car	Road	Vegetation	Sky	Pedestrian	Bicycle	mIoU(%)
SwiftNet (RGB)	83.0	73.4	91.6	96.7	**94.5**	84.7	36.1	82.5	80.3
SwiftNet (Pola)	74.0	66.6	87.1	94.7	91.1	76.1	32.9	65.5	73.5
NLFNet (Ours)	**85.4**	**77.1**	**93.5**	**97.7**	93.2	**85.9**	**56.9**	**85.5**	**84.4**

TABLE III

COMPARATIVE RESULTS (%) ON THE RGB-THERMAL VALIDATION SET [9].

Network	Method	Unlabeled	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU(%)	FPS
DANet [16]	RGB	96.3	71.3	48.1	51.8	30.2	18.2	0.7	30.3	18.8	41.3	-
ERFNet [23]	RGB	95.8	64.8	36.5	42.4	20.5	10.0	0.0	0.0	28.8	33.2	172.1
DUC [40]	RGB	97.7	82.5	69.4	58.9	40.1	20.9	3.4	42.1	40.9	50.7	81.9
HRNet [41]	RGB	98.0	86.9	67.3	59.2	35.3	23.1	1.7	**46.6**	47.3	51.7	-
ACNet [25]	RGB-D	96.7	79.4	64.7	52.7	32.9	28.4	0.8	16.9	44.4	46.3	-
SA-Gate [5]	RGB-D	96.8	73.8	59.2	51.3	38.4	19.3	0.0	24.5	48.8	45.8	-
LDFNet [42]	RGB-D	95.3	67.9	58.2	37.2	30.4	20.1	0.8	27.1	46.0	42.5	-
FuseNet [27]	RGB-D	51.8	75.6	66.3	51.9	37.8	15.0	0.0	21.4	45.0	45.6	**255.3**
RTFNet [43]	RGB-T	**98.5**	87.4	**70.3**	62.7	45.3	**29.8**	0.0	29.1	**55.7**	53.2	34.1
MFNet [9]	RGB-T	96.9	65.9	58.9	42.9	29.9	9.9	0.0	25.2	27.7	39.7	229.9
PSTNet [33]	RGB-T	97.0	76.8	52.6	55.3	29.6	25.1	**15.1**	39.4	45.0	48.4	-
NLFNet (Ours)	RGB-O	97.3	**88.5**	69.0	**63.9**	**47.8**	25.6	6.1	45.0	44.7	**54.3**	35.6

RGB-P-Stack method, we concatenate the RGB image and the polarization image to form a 4-channel image and input it to a single branch to obtain an mIoU of 80.2%, which proves that the polarization image can provide diversified features and serve as supplementary information for the RGB image. However, this method has a lower performance compared with our NLFNet, because the distributions of RGB information and polarization information are different. Pure concatenation will cause interference between them, which will adversely affect the fusion of features of different modalities. NLFNet inputs the RGB image and the polarization image to different paths and uses element-wise addition of the two kinds of information in the final feature aggregation, which obtains an mIoU of 84.4%, indicating that our network model can effectively extract and fuse informative RGB features and polarization features, significantly improving the accuracy of RGB-P semantic segmentation.

Quantitative Performance Study. We verify the performance of NLFNet compared with the SwiftNet baselines on the ZJU-RGB-P dataset, and the quantitative results are shown in Table II. It can be seen that the network model that combines polarization information and RGB information promotes the segmentation of objects with special polarization characteristics, such as *glass* (73.4% to 77.1%) and *cars* (91.6% to 93.5%). In addition, the IoU of other types of objects have also been improved, such as *pedestrians* (36.2% to 56.9%). At the same time, mIoU increases from 80.3% to 84.4%. It shows that our network effectively combines RGB characteristics and polarization characteristics and improves the segmentation accuracy of the network.

TABLE IV

COMPARATIVE RESULTS (%) ON THE DAYTIME AND NIGHTTIME SCENARIOS ON THE RGB-THERMAL VALIDATION SET [9].

Network	Multimodal	Daytime	Nighttime	mIoU(%)
U-Net [4]	✗	✓	✗	37.5
		✗	✓	37.0
SegNet [11]	✗	✓	✗	29.5
		✗	✓	27.4
MFNet [9]	✓	✓	✗	36.1
		✗	✓	36.8
FuseNet [44]	✓	✓	✗	41.0
		✗	✓	43.9
RTFNet [43]	✓	✓	✗	45.8
		✗	✓	**54.8**
FuseSeg [44]	✓	✓	✗	47.8
		✗	✓	54.6
NLFNet	✓	✓	✗	**50.3**
		✗	✓	**54.8**

Table III shows the comparison of the numerical performance of different single-modal and multi-modal networks on the RGB-Thermal dataset provided by the work of MFNet [9]. The compared state-of-the-art networks cover DANet [16], ERFNet [23], and DUC [40], designed for RGB semantic segmentation. We also compare with RGB-D networks including ACNet [25], SA-Gate [5], LDFNet [42], and FuseNet [27], as well as RGB-T networks including RTFNet [43] and MFNet [9]. Our NLFNet inputs RGB information and infrared information into the network for fusion, which attains an mIoU of 54.3%. It achieves good segmentation accuracy for most categories, such as attaining the IoU of *cars* and *bike* with 88.5% and 63.9%, respectively. While our method is inferior to RTFNet in the classes of *Person*, *Car Stop*, and *Bump*, RTFNet relies on a large backbone of ResNet-152, which is significantly more computation-demanding than ResNet-18 used in our approach.

At night, mIoU reaches 54.8% as shown in Table IV,

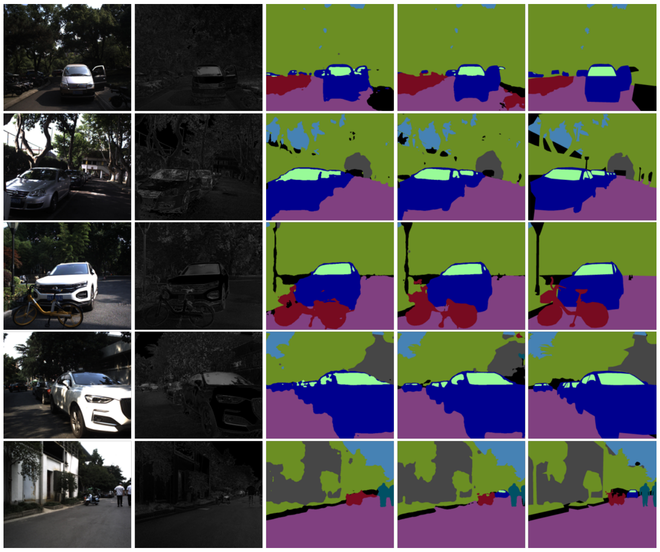

Fig. 8. Qualitative result comparison between the RGB-only network and our RGB-P NLFNet on ZJU-RGB-P dataset [8]. From left to right: RGB image, polarization image (AOLP), RGB-only segmentation, the result of NLFNet, and ground truth.

which demonstrates that NLFNet can effectively use the infrared information of objects to improve the segmentation accuracy in weak-illumination scenarios. The last column of Table III indicates the inference speed for images with a resolution of 640 × 480 on the GTX 1080Ti GPU processor. NLFNet has reached 35.6FPS (Frames Per Second). This shows that NLFNet still has a good real-time performance while improving the accuracy of image segmentation.

In Table V, we further compare NLFNet with some representative RGB or RGB-D networks on the Cityscapes validation set. NLFNet achieves an mIoU of 72.3%, which proves that the position and contour information in the depth image can supplement RGB features, and our method can effectively fuse multimodal features to achieve good segmentation accuracy in urban driving scenes.

TABLE V

COMPARISON OF SEMANTIC SEGMENTATION METHODS ON THE VALIDATION SET OF CITYSCAPES [7].

Network	Multimodal	mIoU(%)
FCN8s [2]	✗	65.3
DeepLabV2-CRF [13]	✗	70.4
ERFNet [23]	✗	65.8
ERF-PSPNet [45]	✗	64.1
SwiftNet [24]	✗	72.0
VGG-D (ScaleInvariant) [46]	✓	64.4
LDFNet [42]	✓	68.5
GoogLeNet (NiN-2) [47]	✓	69.1
RFBNet (ERFNetEnc) [48]	✓	72.0
NLFNet (Ours)	✓	**72.3**

Qualitative Performance Study. In order to obtain qualitative results, we use SwiftNet based on RGB information and NLFNet based on multimodal input information to conduct experiments on the ZJU-RGB-P, RGB-Thermal, and Cityscapes datasets, respectively, and the visual comparison results are shown in Fig. 8, Fig. 9, and Fig. 10. In Fig. 8, NLFNet, which incorporates polarization images, has better segmentation results for objects with special polarization characteristics, such as *cars*, *bicycles*, and *glasses* like the car windows.

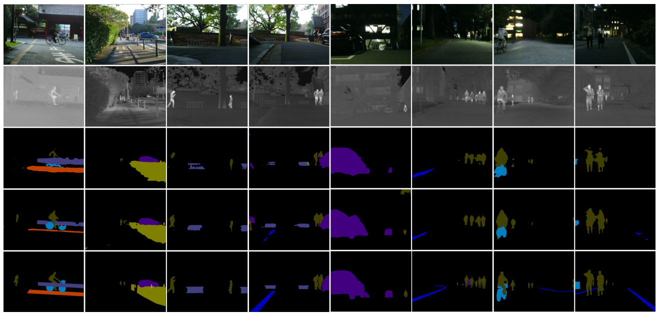

Fig. 9. Qualitative result comparison between the RGB-only network and our RGB-T NLFNet on RGB-Thermal dataset [9]. From top to bottom: RGB image, thermal infrared image, RGB-only segmentation, the result of NLFNet, and ground truth.

NLFNet integrates the special infrared thermal information of the objects, which significantly improves the segmentation accuracy of *pedestrians* in the nighttime environment, as shown in Fig. 9. In Fig. 10, NLFNet, which incorporates depth features, can make good use of the position and contour information of objects, improving the segmentation accuracy of *cars*, *trucks*, and *pedestrians*. The comprehensive analysis with these results proves that our NLFNet can effectively integrate the features coming from different modalities and improve the accuracy of segmentation, which has a high generalization capacity across various sensor combinations.

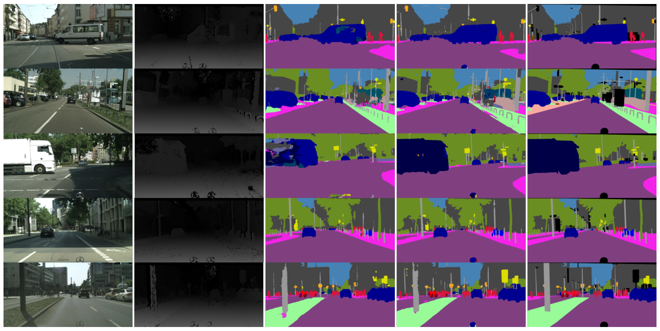

Fig. 10. Qualitative result comparison between the RGB-only network and our RGB-D NLFNet on Cityscapes dataset [7]. From left to right: RGB image, depth image, RGB-only segmentation, the result of NLFNet, and ground truth.

V. CONCLUSION

In this work, we investigate generalizable multimodal perception and propose a semantic segmentation network NLFNet suitable for outdoor scene understanding, which effectively solves the problem of object segmentation in various challenging scenarios. The designed NLF module can perform adaptive feature extraction and fusion of complementary information from different modal input images, and leverage the dependence information with long-range contextual and positional priors to improve the accuracy of semantic segmentation. We have conducted extensive experiments and analysis on ZJU-RGB-P, RGB-Thermal (MFNet), and RGB-D (Cityscapes) datasets, which verify the effectiveness and generalization ability of NLFNet across several multimodal sensor combinations.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.

[2] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015.

[5] X. Chen *et al.*, “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation,” in *ECCV*, 2020.

[6] J. Zhang, K. Yang, and R. Stiefelhagen, “ISSAFE: Improving semantic segmentation in accidents by fusing event-based data,” in *IROS*, 2021.

[7] M. Cordts *et al.*, “The cityscapes dataset for semantic urban scene understanding,” in *CVPR*, 2016.

[8] K. Xiang, K. Yang, and K. Wang, “Polarization-driven semantic segmentation via efficient attention-bridged fusion,” *OE*, 2021.

[9] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” in *IROS*, 2017.

[10] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, “Deep multimodal fusion for semantic image segmentation: A survey,” *IVC*, 2020.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *TPAMI*, 2017.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *TPAMI*, 2018.

[14] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018.

[15] S. Choi, J. T. Kim, and J. Choo, “Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks,” in *CVPR*, 2020.

[16] J. Fu *et al.*, “Dual attention network for scene segmentation,” in *CVPR*, 2019.

[17] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen, “Capturing omni-range context for omnidirectional segmentation,” in *CVPR*, 2021.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *CVPR*, 2018.

[19] S. Zheng *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *CVPR*, 2021.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” *arXiv*, 2021.

[21] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, “Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world,” in *ICCVW*, 2021.

[22] B. Cheng, A. G. Schwing, and A. Kirillov, “Per-pixel classification is not all you need for semantic segmentation,” *arXiv*, 2021.

[23] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, “ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation,” *T-ITS*, 2018.

[24] 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 *CVPR*, 2019.

[25] X. Hu, K. Yang, L. Fei, and K. Wang, “ACNet: Attention based network to exploit complementary features for rgbd semantic segmentation,” in *ICIP*, 2019.

[26] 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,” *RA-L*, 2020.

[27] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture,” in *ACCV*, 2016.

[28] J. Vertens, J. Zürn, and W. Burgard, “HeatNet: Bridging the day-night domain gap in semantic segmentation with thermal images,” in *IROS*, 2020.

[29] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, “Bridging the day and night domain gap for semantic segmentation,” in *IV*, 2019.

[30] L. Sun, K. Wang, K. Yang, and K. Xiang, “See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion,” in *SPIE*, 2019.

[31] H. Zhou, L. Qi, Z. Wan, H. Huang, and X. Yang, “RGB-D co-attention network for semantic segmentation,” in *ACCV*, 2020.

[32] G. Zhang, J.-H. Xue, P. Xie, S. Yang, and G. Wang, “Non-local aggregation for RGB-D semantic segmentation,” *SPL*, 2021.

[33] 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 *ICRA*, 2020.

[34] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon, “MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation,” *RA-L*, 2021.

[35] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, “Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation,” *TNNLS*, 2021.

[36] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, “ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation,” in *CVPR*, 2021.

[37] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in *CVPR*, 2005.

[38] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *IJCV*, 2015.

[39] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

[40] P. Wang *et al.*, “Understanding convolution for semantic segmentation,” in *WACV*, 2018.

[41] J. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *TPAMI*, 2020.

[42] S.-W. Hung, S.-Y. Lo, and H.-M. Hang, “Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation,” in *ICIP*, 2019.

[43] Y. Sun, W. Zuo, and M. Liu, “RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes,” *RA-L*, 2019.

[44] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, “FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion,” *T-ASE*, 2020.

[45] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, “Unifying terrain awareness through real-time semantic segmentation,” in *IV*, 2018.

[46] I. Kreso, D. Causevic, J. Krapac, and S. Segvic, “Convolutional scale invariance for semantic segmentation,” in *GCPR*, 2016.

[47] L. Schneider *et al.*, “Multimodal neural networks: RGB-D for semantic segmentation and object detection,” in *SCIA*, 2017.

[48] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, “RFBNet: Deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation,” *arXiv*, 2019.