2023 IEEE International Conference on Robotics and Biomimetics (ROBIO) | 979-8-3503-2570-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ROBIO58561.2023.10354613

IGFNet: Illumination-Guided Fusion Network for Semantic Scene Understanding using RGB-Thermal Images

Haotian Li[1] and Yuxiang Sun[2,*]

***Abstract*— Semantic scene understanding is a fundamental task for autonomous driving. It serves as a build block for many downstream tasks. Under challenging illumination conditions, thermal images can provide complementary information for RGB images. Many multi-modal fusion networks have been proposed using RGB-Thermal data for semantic scene understanding. However, current state-of-the-art methods simply use networks to fuse features on multi-modality inexplicably, rather than designing a fusion method based on the intrinsic characteristics of RGB images and thermal images. To address this issue, we propose IGFNet, an illumination-guided fusion network for RGB-Thermal semantic scene understanding, which utilizes a weight mask generated by the illumination estimation module to weight the RGB and thermal feature maps at different stages. Experimental results show that our network outperforms the state-of-the-art methods on the MFNet dataset. Our code is available at: https://github.com/lab-sun/IGFNet.**

I. Introduction

Semantic segmentation aims to label input images into pixel-wise semantic classification maps. It is a fundamental technology to understand scenes in many applications [1], [2]. In robotic-related applications, semantic segmentation can be used for many tasks, such as road detection [3], trajectory prediction [4], negative obstacles segmentation [5], [6], and decision making [7].

Due to the rapid development of deep learning, semantic segmentation using RGB images has made great progress [8], [9], [10]. However, RGB cameras can only capture high-quality images with rich texture information when the environment lighting conditions are satisfactory. This will lead to the failure of semantic segmentation model based on RGB images when encountering poor exposure conditions, such as overexposure of the sky during the day, underexposure and glare at night, etc. Therefore, thermal images are used as complementary information to the RGB images under the aforementioned challenging illumination conditions [11]. Due to the similarity between thermal images and RGB images, RGB-Thermal semantic segmentation has also developed rapidly in recent years [12], [13], [14].

Though methods based on convolutional neural network (CNN), such as MFNet [14] and RTFNet [15], have achieved acceptable segmentation performance, they still cannot meet the usage requirements. In contrast to CNN-based semantic segmentation methods [14], [15], [16], transformer-based methods [17], [18], [19], [20] establish long-range contextual dependencies between pixels of images by treating them as sequences using the transformer. Building upon Segformer [21], CMX [22] employs transformers to extract features for the task of RGB-Thermal semantic segmentation. However, whether it is a method based on CNN or a method based on vision transformer, the fusion of RGB and thermal images does not consider the intrinsic characteristics of different data, but design a simple inexplicably fusion module to use the network as a black box. In fact, RGB images can provide rich texture information that thermal images do not have, whereas thermal images may be affected by strong heat sources in the scene. However, this information will be lost in overexposed areas at night (e.g., glare areas caused by vehicle headlights), underexposed areas at night (e.g., areas lacking ambient light) and overexposed areas during the day (e.g., sky) [23].

To address this issue, we have developed a novel Illumination-Guided Fusion Network (IGFNet) that incorporates an Illumination Estimation Module (IEM) and Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM). The IEM employs a Gaussian filter to map the grayscale value of each pixel in the RGB image to a weight factor, representing the significance of the RGB information of that pixel during fusion. This allows the multi-modal fusion to be guided by the illumination of the environment. Subsequently, the IGCM-RM utilizes the weight mask generated by the IEM to recalibrate the multi-modal feature maps at various stages. This enables the feature maps to be fused according to the significance of each pixel. Our main contributions can be summarized as follows:

1) Rethink the fusion of RGB and thermal images, proposing an Illumination Estimation Module (IEM) to generate a weight mask as a clue.
2) Propose an Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM), using the clue to recalibrate the feature maps of RGB and thermal images.
3) Propose a novel RGB-Thermal semantic segmentation network named IGFNet with IEM and IGCM-RM.
4) Our proposed method demonstrates state-of-the-art performances on the MFNet dataset [14] and our code is open sourced.

II. Related Work

A. CNN-based RGB-Thermal Semantic Segmentation

Current research on RGB-Thermal semantic segmentation primarily focuses on developing delicate models to efficiently fuse multi-modal data, thereby enhancing segmentation performance. MFNet [14] and RTFNet [15] both utilized the

[1]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: haotian.li@connect.polyu.hk).

[2]City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

*Corresponding author: Yuxiang Sun.

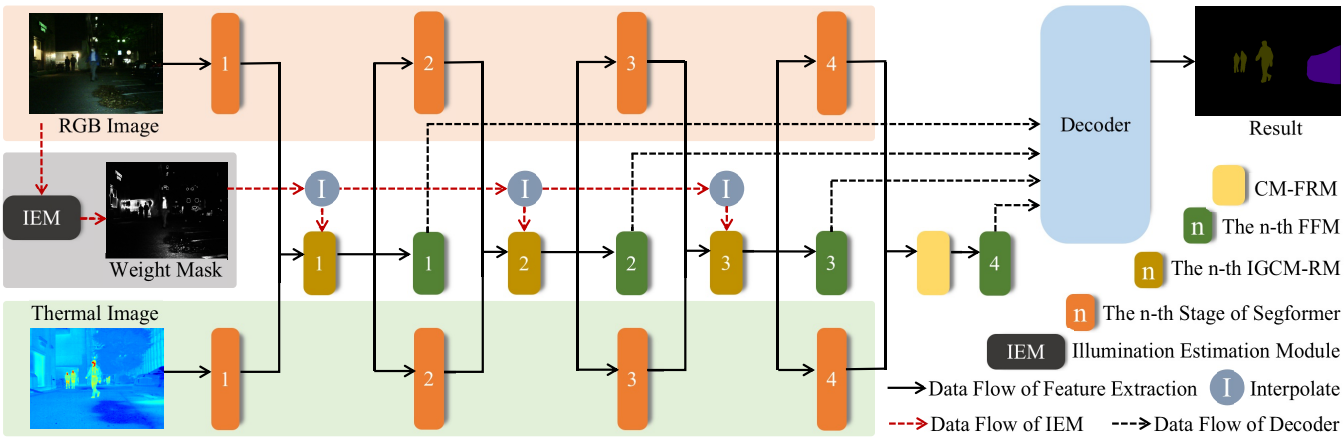

Fig. 1: The overview of IGFNet for RGB-Thermal semantic segmentation. The inputs are an RGB image and a thermal image. IGFNet consists of a four-stage RGB encoder, a four-stage thermal encoder, a Illumination Estimation Module (IEM), a three-stage llumination-Guided-Cross-Modal Rectification Module (IGCM-RM), a Cross-Modal Feature Rectification Module (CM-FRM), a four-stage Feature Fusion Module (FFM) and a decoder. The encoder and decoder are borrowed from Segformer [21]. The CM-FRM and FFM are borrowed from CMX[22]. The thermal image is colored with the *jet* color map.

feature fusion structure [22] (i.e., structure using two encoders and one decoder) to fuse the feature maps. Sun et al. [24] proposed FuseSeg, which continued the previous structure, and obtaind better segmentation results through a two-stage fusion strategy. Xu et al. [25] proposed AFNet based on the attention fusion module, utilizing attention to guide the fusion of RGB and thermal features. To minimize the modality differences between RGB and thermal features, ABMDRNet [11] employs a bridging-then-fusing strategy that utilizes a bi-directional image-to-image translation-based method. Zhou et al. [26] proposed GMNet that divides the feature extraction into three levels: junior, intermediate, and senior. It is evident that the majority of the above methods focus on developing advanced models that utilize the complementary information from different modalities to enhance the accuracy of segmentation results.

B. Transformer-based RGB-Thermal Semantic Segmentation

Unlike CNN-based methods that employ channel or spatial attention to improve performance, Vaswani et al. [20] proposed an attention mechanism that only relies on self-attention to establish long-range dependencies among the input. DANet [27] employed self-attention to selectively build connections between local features and global dependencies. Meanwhile, CCNet [28] proposed criss-cross attention module to obtain the dense and global contextual information. Further on, Dosovitskiy et al. [17] and Touvron [18] proposed Vision Transformer (ViT) and a teacher-student strategy based purely on attention, respectively. Building on this, SETR [29] processed an input image as a sequence of patches and employed a pure transformer structure to extract global context at different stages. Similarly, Segmenter [30], designed based on ViT, achieves semantic segmentation using a purely transformer-based approach. Segformer [21] introduced a hierarchical transformer structure to capture multi-resolution features and combined these features to predict semantic labels using a decoder composed of multilayer perceptrons (MLPs). Zhang et al. [22] introduced CMX, an extension of Segformer to multimodal tasks, which achieves semantic segmentation of RGB-Thermal images.

Though the aforementioned methods have continuously improved the segmentation capabilities through a series of advanced models, they only treat these models as black boxes, allowing them to learn the connections between different modalities on their own, resulting in a lack of interpretability. Different from previous works, we rethink the intrinsic characteristics of RGB-Thermal images, proposing an fusion network guided by illumination.

III. The Proposed Method

A. Network Overview

The overview of IGFNet is shown in Fig. 1. As illustrated in Fig. 1., we adopt the feature fusion structure comprising two encoders and one decoder. The two encoders are responsible for extracting features from the RGB and thermal images, respectively.

In poorly exposed environments, thermal images can provide complementary information to RGB images. By combining the information from both modalities, it is possible to achieve mutual correction and improve segmentation accuracy. Specifically, in extreme environments such as glare areas caused by vehicle headlights or areas with low ambient light at night, RGB images may lose texture information. In these cases, thermal images can provide complementary features to better understand the scene.

To exploit this property, we proposed the Illumination Estimation Module (IEM) to guide multi-modal fusion by generating an illumination-based weight mask. Subsequently, we used interpolation modules to hierarchical downsample the weight mask, and the results were applied to the following feature fusion modules. Then, we proposed the Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM) to utilize the weight mask in guiding the rectification of both RGB and thermal features. In this manner, the rectified

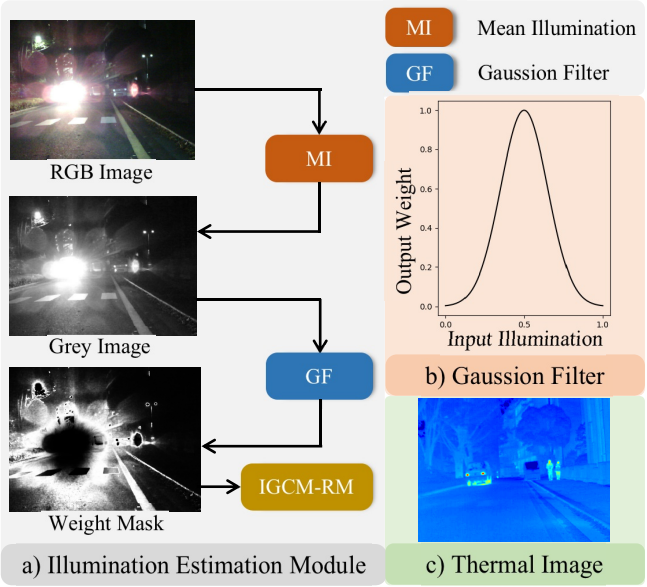

Fig. 2: a) The pipeline of our proposed IEM. The input to IEM is an RGB image, from which we derive a grey image by calculating the mean illumination. We then obtain the weight mask by applying a Gaussian filter to this grey image. b) The graph illustrates the application of the Gaussian filter, mapping the grey image to the weight mask. c) The corresponding thermal image shows complementary information.

features are divided into two streams, one stream is used as the input of the next stage encoder, while the other stream is used as input to the Feature Fusion Module (FFM). Since the weight mask essentially represents spatial attention, and the spatial information in high-level features decreases as the network depth increases, we only apply the IGCM-RM to the first three stages of the encoder. Then, we apply Cross-Modal Feature Rectification Module (CM-FRM) to the forth stage of the encoder. This will be further demonstrated in Section IV-C.

B. Illumination Estimation Module

As previously discussed, RGB and thermal images provide distinct information under varying illumination conditions. To leverage this, we have developed the IEM to generate a weight mask indicating the importance of each RGB pixel based on illumination.

The pipeline of the IEM is shown in the part a) of Fig. 2. The first step in our process involves calculating the average grayscale value of the three channels in the RGB image. This produces a grey image that reflects the illuminance values at different positions. Then, we propose a Gaussian filter to map the grayscale values of individual pixels to weights representing the importance of RGB information. The Gaussian function of this filter is described as follows:

$$f(x) = exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

where $\mu = 0.5$ and $\sigma = 0.15$. As shown in the part b) of Fig. 2, μ determines the midpoint position of the Gaussian function. We have chosen a value of $\mu = 0.5$, as this yields the best exposure and most abundant texture information in the RGB image. At this point, the corresponding output weight is 1, indicating that the RGB information is the most important. For instance, the RGB image in Fig. 2 shows the glare-affected area caused by vehicle headlights, and we cannot find any useful texture in this area. In contrast, the corresponding thermal image provides complementary information. The weight mask we have obtained maps the glare-affected area to 0 and the well-exposed areas to 1 according to the Gaussian distribution. This guides the network to primarily utilize thermal features in this area and RGB features in well-exposed areas during the feature fusion stage. Additionally, we can observe that the edges of the glare-affected area are assigned values close to 1 in the weight mask. This indicates that RGB information will be primarily utilized in feature fusion, facilitating accurate segmentation of object edges.

C. llumination-Guided-Cross-Modal Rectification Module

After obtaining the weight mask from the IEM, we proposed the IGCM-RM to recalibrate the features of RGB and thermal images, guided by the weight mask. The structure of the IGCM-RM is shown in Fig. 3. First, we obtain the inverted mask by subtracting the weight mask from a all-ones matrix. This inverted mask represents the importance of various positions within the thermal feature map during fusion.

For better illustration, we will use the stream of the weight mask $Mask \in \mathbb{R}^{H\times W\times 1}$ and the RGB feature map $F_{RGB} \in \mathbb{R}^{H\times W\times C}$ as an example. To ensure that the dimensions of the weight mask correspond with the feature maps, we expand the weight mask along the channel dimension to get $Mask \in \mathbb{R}^{H\times W\times C}$. Instead of performing an element-wise multiplication of $Mask \in \mathbb{R}^{H\times W\times C}$ and $F_{RGB} \in \mathbb{R}^{H\times W\times C}$, we first concatenate them. Then, the resulting output is fed into a block, which consists of two 1×1 convolution layers, a 3×3 convolution layer, and a ReLU layer. The purpose of this block is to adaptively recalibrate the values of the feature map F_{RGB}, guided by the weight mask $Mask$. The output of this stream undergoes an element-wise addition with the output of another stream. This second stream originates from F_{RGB} and passes through a 1×1 convolution layer, with the goal of providing the original information of the RGB features. The outputs of this stage are $R_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $R_T \in \mathbb{R}^{H\times W\times C}$.

The rectified RGB feature, $R_{RGB} \in \mathbb{R}^{H\times W\times C}$, which is embedded with illumination information, is then concatenated with the corresponding thermal feature, $R_T \in \mathbb{R}^{H\times W\times C}$. Subsequently, max pooling and average pooling are applied to the output features, $R_{RGB-T} \in \mathbb{R}^{H\times W\times 2C}$, respectively. Then, the concatenated output vectors, $Y \in \mathbb{R}^{4C}$, is fed into the MLP layer. The output of this MLP layer contain two channel weights, $C_{RGB} \in \mathbb{R}^{C}$ and $C_T \in \mathbb{R}^{C}$. Similarly, two spatial weights, $S_{RGB} \in \mathbb{R}^{H\times W\times 1}$ and $S_T \in \mathbb{R}^{H\times W\times 1}$, can be generated by another MLP layer. The output of IGCM-RM is formulated as:

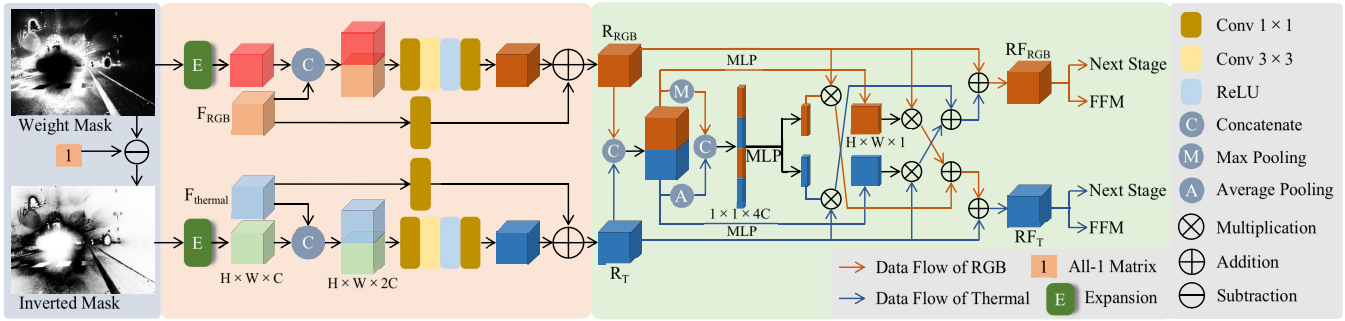

Fig. 3: The structure of our proposed IGCM-RM. The inputs to this module are the feature maps of the RGB and thermal images from the encoder, as well as the weight mask generated by the IEM. After rectification guided by the weight mask, the rectified feature maps are divided into two streams. One stream serves as input to the subsequent stage encoder, while the other is input to the FFM for feature fusion. The part in green background is borrowed from CMX[22].

$$RF_{RGB} = R_{RGB} + \lambda_1 C_{RGB} \otimes R_{RGB} + \lambda_2 S_{RGB} \odot R_{RGB}$$

$$RF_T = R_T + \lambda_1 C_T \otimes R_T + \lambda_2 S_T \odot R_T \quad (2)$$

where $\otimes$ denotes channel-wise multiplication, $\odot$ denotes spatial-wise multiplication, λ_1 and λ_2 are two hyper-parameters. Inspired by [22], we set $\lambda_1 = \lambda_2 = 0.5$. Finally, the rectified features, RF_{RGB} and RF_T, will be sent to the next stage encoder and the FFM for feature fusion.

IV. Experimental Results and Discussions

A. Dataset

For our experiments, we utilized the available MFNet dataset released in [14]. This dataset contains 9 classes (i.e., unlabelled background, car, person, bike, curve, car stop, guardrail, color cone, and bump), with manually annotated labels for semantic segmentation. The dataset contains 2,390 pairs of RGB and thermal images, derived from the original dataset consisting of 1,596 pairs. We employed the same split scheme as described in [14] to train our network, with 1568 pairs used for training, 392 for validation, and 393 for testing.

B. Implementation Details

Our IGFNet is implemented using PyTorch and trained and tested on a PC equipped with an NVIDIA RTX 3090 (24GB RAM). For the encoder, we utilize the Mix Transformer encoder (MiT-B2) pretrained on ImageNet [31] as the backbone, while the decoder employs an MLP, both of which were proposed in SegFormer [21]. We train the network using the AdamW optimizer [32], with a weight decay set to 0.01. The initial learning rate is set to $6e^{-5}$, and a poly learning rate schedule is employed. The momentum and the decay strategy are set to 0.9. The batch size is set to 4 during training and we use cross-entropy as the loss function. We use Precision (Pre), Accuracy (Acc), F1, Intersection over Union (IoU), Floating Point Operations (FLOPs) and Parameters (Params) to evaluate our network in Section IV-C, while using Acc and IoU in Section IV-D.

TABLE I: The results of the ablation study. '✓' means the IEM and IGCM-RM are used behind the n-th stage of the encoder. '—' means the CM-FRM [22] is used hehind the n-th stage of the encoder. The best results are hightlighted in bold font.

No.	Stage				mPre	mAcc	mF1	mIoU	FLOPs(G)	Params(M)
	1st	2nd	3rd	4th						
(A)	—	—	—	—	**75.93**	67.32	69.85	58.19	**66.87**	**66.57**
(B)	✓	—	—	—	73.25	72.67	70.39	58.46	67.54	66.60
(C)	—	✓	—	—	71.65	72.01	69.97	58.40	67.52	66.68
(D)	—	—	✓	—	73.88	71.36	69.96	58.24	67.86	67.29
(E)	—	—	—	✓	72.30	72.07	69.73	58.18	67.50	68.42
(F)	✓	✓	—	—	74.50	72.04	70.77	58.89	68.19	66.71
(G)	✓	✓	✓	—	73.92	**72.91**	**71.04**	**59.01**	69.18	67.44
(H)	✓	✓	✓	✓	73.03	72.18	70.21	58.47	69.82	69.29

C. Ablation Study

We conducted an ablation study to evaluate the benefits of IEM and IGCM-RM, and subsequently selected the optimal structure for our IGFNet. In this section, we use CMX [22] based on MiT-B2 as the baseline. In the first part of our study, we positioned the two modules (IEM and IGCM-RM) behind the encoder at different stages. For instance, in No.B network, the two modules were placed behind the first stage encoder. Subsequently, the rectified RGB and thermal features were input into the FFM and the second stage encoder. Table I shows that, except for No.E, the segmentation performance (i.e., mF1 and mIoU) of all networks (i.e., No.B, No.C, and No.D) with IEM and IGCM-RM added is better than the baseline. Additionally, our results indicate that the network's performance gradually decreases as the stage at which the two modules are inserted is delayed. This can be attributed to the fact that the weight mask obtained through IEM essentially represents spatial attention. As the stage of the encoder progresses, the spatial information contained within the high-level features learned by the network diminishes. Hence, we designed several combinations based on this characteristic in the second stage. For instance, network No.F placed IEM and IGCM-RM behind the first and second stages at the same time. As shown in Table I, the network No.G obtains the best performance. In contrast, the performance of network No.H was found to

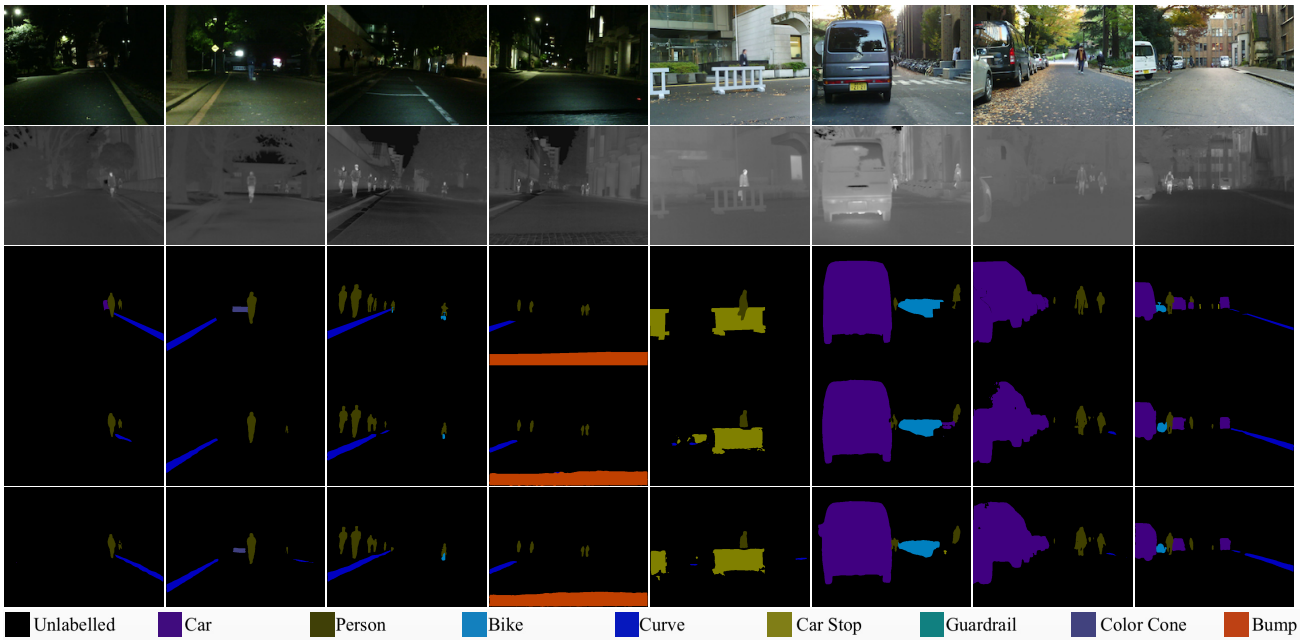

Fig. 4: Sample qualitative demonstrations. The rows from top to bottom are RGB images, thermal images, ground truth, CMX results, and IGFNet results.

TABLE II: The comparative per-class results on the MFNet dataset [14]. To evaluate the methods, we utilized the IoU of each class and the mean IoU of all classes. The results demonstrate the superiority of our IGFNet, with the top two results in each column are highlighted in red and blue.

Method	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [14]	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	0.0	30.3	25.2	30.0	27.7	45.1	39.7
RTFNet [15]	93.0	87.4	79.3	70.3	**76.8**	**62.7**	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	**74.7**	55.7	63.1	53.2
AFNet [25]	91.2	86.0	76.3	67.4	72.8	62.0	49.8	43.0	35.3	28.9	24.5	4.6	50.1	44.9	61.0	56.6	62.2	54.6
ABMDRNet [11]	**94.3**	84.8	**90.0**	69.6	75.7	60.3	64.0	45.1	**44.1**	**33.1**	31.0	5.1	**61.7**	47.4	66.2	50.0	69.5	54.8
FEANet [16]	**93.3**	87.8	82.7	71.1	**76.7**	61.1	**65.5**	46.5	26.6	22.1	**70.8**	6.6	**66.6**	**55.3**	**77.3**	48.9	**73.2**	55.3
CMX(MiT-B2) [22]	92.2	**89.4**	81.3	**74.8**	73.4	**64.7**	63.5	**47.3**	38.8	30.1	36.3	**8.1**	53.3	**52.4**	67.7	**59.4**	67.3	**58.2**
IGFNet(ours)	93.2	**88.0**	**83.4**	**74.0**	71.8	**62.7**	**67.6**	**48.2**	**45.4**	**36.0**	**68.5**	**14.2**	58.8	**52.4**	68.3	**57.5**	**72.9**	**59.0**

be inferior to that of the network No.G. This observation is in alignment with our conclusion before. According to FLOPs and Params, it can be seen that adding IEM and IGCM-RM does not bring much computational complexity to the framework. Based on the results of our ablation study, we choose network No.G as the optimal structure of our IGFNet.

D. Comparative Study

We compared our proposed IGFNet with the state-of-the art methods, including MFNet [14], RTFNet [15], AFNet [25], ABMDRNet [11], FEANet [16], and CMX(MiT-B2) [22].

1) Quantitative Results: As shown in Table II, we can find that the mIoU of our proposed IGFNet is higher than other methods. Specifically, our IGFNet demonstrates a significant advantage over other methods in accurately segmenting the most challenging categories of car stop and guardrail. Additionally, it achieves the highest mIoU for the segmentation of curve. This shows that IGFNet significantly improves the segmentation performance of small and difficult-to-recognize objects by using illumination-guided feature fusion. Though FEANet achieves the highest Acc in the segmentation of guardrail and bump, its corresponding IoU do not show a corresponding performance. In contrast, our proposed IGFNet consistently maintains a highly competitive performance of both Acc and IoU across the segmentation of all classes.

2) Qualitative Results: To intuitively demonstrate the effectiveness of our IGFNet, we selected four groups of daytime and nighttime images as examples, respectively. We compared them against the two best-performing methods, i.e., CMX(MiT-B2) and IGFNet, listed in Table II. As shown in Fig. 4., the segmentation performance of our IGFNet on curve and car stop is significantly superior to that of CMX. This is consistent with the quantitative results in the first part of Section IV-D, which demonstrate that the strategy of utilizing illumination to guide feature map fusion in IGFNet can effectively enhance segmentation accuracy, especially for objects that are difficult to distinguish.

V. CONCLUSIONS

We proposed here a novel semantic segmentation network IGFNet, which introduces an interpretable RGB-Thermal fusion network and utilizes illumination to guide the fusion of multi-modal features. The IEM employs a Gaussian filter to generate a weight mask, indicating the significance of each RGB pixel based on the illumination. Then, the generated

weight mask is utilized as a cue within the IGCM-RM to rectify the multi-modal features, enabling their more effective fusion. The experimental results show that our proposed IGFNet achieves better performance than the state-of-the-art methods. However, there are still some limitations in our IGFNet. For instance, we found that the semantic segmentation results of consecutive frames lack consistency, which can lead to hesitation or misjudgment in autonomous driving decisions. Therefore, we would like to solve this issue in the future work.

ACKNOWLEDGEMENT

This work was supported by the Hong Kong Innovation and Technology Fund under Grant ITS/145/21.

REFERENCES

[1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1760–1770.

[3] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 750–10 760, 2022.

[4] Y. Sun, W. Zuo, and M. Liu, "See the future: A semantic segmentation network predicting ego-vehicle trajectory with a single monocular camera," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3066–3073, 2020.

[5] Z. Feng, Y. Feng, Y. Guo, and Y. Sun, "Adaptive-mask fusion network for segmentation of drivable road and negative obstacle with untrustworthy features," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–6.

[6] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "Inconseg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4871–4878, 2023.

[7] Y. Feng, W. Hua, and Y. Sun, "Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9780–9791, 2023.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2633–2642.

[12] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1223–1235, 2022.

[13] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for rgb-t semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[15] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[16] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4467–4473.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[22] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *arXiv preprint arXiv:2203.04838*, 2022.

[23] K. Song, Y. Zhao, L. Huang, Y. Yan, and Q. Meng, "Rgb-t image analysis technology and application: A survey," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105919, 2023.

[24] 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 Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2020.

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

[26] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[28] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[29] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[30] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.