IAFFNet: Illumination-Aware Feature Fusion Network for All-Day RGB-Thermal Semantic Segmentation of Road Scenes

YA-LI HOU¹, (Member, IEEE), YAN JIA¹, ZHIJIANG HOU², XIAOLI HAO¹, AND YAN SHEN¹, (Member, IEEE)

¹School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
²Library, Tianjin University of Technology, Tianjin 300384, China

Corresponding author: Ya-Li Hou (ylhou@bjtu.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant 4202061; in part by the National Natural Science Foundation of China under Grant 61771042, Grant 71804123, and Grant 61702032; and in part by the Fundamental Research Funds for the Central Universities of China under Grant 2018JBM007.

ABSTRACT Semantic segmentation based on RGB and thermal images is an effective way to achieve an all-day understanding of road scenes. However, how to fuse the RGB and thermal information effectively remains an open problem. By studying fusion strategies at different stages, an illumination-aware feature fusion network has been developed for all-day semantic segmentation of urban road scenes in this paper, called IAFFNet. At the encoding stage, we introduce a bi-directional guided feature fusion module to effectively recalibrate and unify both RGB and thermal information. At the decoding stage, we have developed adaptive fusion modules to fuse low-level details and high-level semantic information. Finally, we have developed a decision-level illumination-aware strategy to achieve robust all-day segmentation. As far as we know, we are the first to incorporate illumination clues explicitly into RGB-T semantic segmentation. Extensive experimental evaluations demonstrate that the developed method can achieve remarkable performance on public datasets compared with state-of-the-art methods. The mIoU of all-day on a public RGB-thermal urban scene dataset has achieved 56.6%.

INDEX TERMS Semantic segmentation, thermal images, road scenes, information fusion.

I. INTRODUCTION

As an effective approach to scene understanding, semantic segmentation of urban scenes has become a basic component of autonomous driving. Numerous studies confirm that semantic segmentation based on the deep neural networks show a remarkable performance, especially the convolutional neural networks (CNNs) [1], [2], [3], [4], [5]. However, most semantic segmentation methods are based on visible RGB images, which heavily rely on lighting and weather conditions. Thermal cameras are proved to be useful to help achieve all-day semantic segmentation of road scenes in [6].

Visible cameras work in the visible light spectrum, ranging from 0.4 µm to 0.7 µm. Under good lighting conditions, RGB images can usually describe more details for the appearances of objects. Thermal cameras can detect wavelengths up to 14 µm, and they can see almost everything that creates heat. Even under poor lighting conditions, objects with thermal radiation tend to show significant clues. However, thermal images usually have low resolutions and contain a lot of noise because of their imaging principles [7]. Although some previous works have shown the effectiveness of using both RGB and thermal images [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], how to fuse the two modalities remains an open problem.

Based on a careful study of fusion strategies at different stages, we have developed an illumination-aware feature fusion
fusion network (IAFFNet) for RGB-T semantic segmentation in this paper. The developed method has obtained improved results against state-of-the-art frameworks. The main contributions of the developed IAFFNet are as follows:

- We introduce a bi-directional guided feature fusion module to fully-exploit the information from both RGB and thermal images at the encoding stage.
- We develop adaptive fusion strategies to fuse feature maps from the encoder with upsampled features in the decoder.
- Illumination-aware fusion strategies are proposed to achieve robust all-day semantic segmentation based on RGB and thermal images. To our knowledge, we are the first to incorporate illumination information into RGB-T semantic segmentation explicitly.

II. RELATED WORK

Early work on RGB-T semantic segmentation focuses on network architectures, in which simple operations such as element-wise addition and concatenation are used to integrate the two modalities [6], [8], [9], [10]. Based on an encoder-decoder architecture, Ha et al [6] proposed a real-time semantic segmentation network using RGB and thermal data as a two-stream encoding branch in 2017. The results demonstrated that the dual-branch architecture outperforms the method with thermal information as an additional input channel. In RTFNet, Sun et al [8] also designed two parallel encoders and one decoding branch. At the encoding stage, they used element-wise summation to gradually integrate the information of the thermal branch into the RGB encoder branch, and they designed a novel Uception module to form the decoder. In FuseSeg [9], they skip-connected fused features at the encoding stage with the upsampled features at the decoding stage. The skip-connections are implemented with the concatenation of two feature maps. To exploit the large amount of RGB data and annotations better, [10] used an independent RGB stream, and the results are cascaded with the original RGB and thermal images into another stream to get the segmented result. Recently, MMNet [11] developed a multi-modal multi-stage architecture for semantic segmentation. Features are exploited at the first stage, and cross-modal knowledge is fused and refined at the second stage.

Simple fusion methods usually ignore the particularity and differences of the RGB and thermal information. With the popularity of the attention mechanism, adaptive fusion of two modalities has shown its superiority. In [12], Deng et al. used a sequential channel-wise and spatial-wise attention module to enhance RGB and thermal features separately. After that, features from the two modalities are element-wisely summed up and propagated along the RGB branch. In MLFNet [13], RGB and thermal features are passed through the SE module [14] respectively and then added element-wisely. At the decoding stage, multi-level skip connections are used to achieve more comprehensive feature fusion. Both the same-level and low-level features from the encoder are concatenated with the features from the previous decoding layer. In ABMDRNet [15], a bi-directional image-to-image translation method is used to reduce the differences between RGB features and thermal features first, and then the multi-modality features are adaptively fused in a channel-wise way. To exploit the contextual information of the cross-modality features along with the spatial and channel dimensions, two modules are also developed for the high-level features at the encoding stage in [15]. In MFFNet [16], information in the RGB branch is element-wisely summed up to the thermal branch at the encoding stage. At the decoding stage, features from different levels are concatenated and enhanced by a spatial-wise attention module to obtain the final segmentation results. The GMNet [17] proposed by Zhou et al. in 2021 artificially divides multi-layer features into three levels and uses different fusion modules to fuse the features. At the decoding stage, the fused features are integrated with the decoder features by element-wise addition. In [18], features from the RGB and thermal branch are encoded separately at the encoding stage, and fused features are only used for the decoder. Fused features from each level are skip-connected to the decoding stage to produce the segmentation results. In [16], [17], and [18], multi-task supervision strategies are used to enhance the learning ability of the network and improve segmentation accuracy in terms of semantic, binary, and boundary characteristics.

III. PROPOSED METHOD

The developed illumination-aware feature fusion network (IAFFNet) based on RGB and thermal images is shown in FIGURE 1. As it has shown good performance for RGB-T based semantic segmentation, we have used the baseline framework of RTFNet [8] in our method. As in RTFNet, the developed network includes two encoding branches to extract RGB and thermal image features. The two branches are symmetrical except for the first convolution layer because RGB and thermal images contain different channels. The backbones of the encoding branches consist of four Resnet [19] blocks. The decoder follows the design of Uception blocks in RTFNet to restore the resolution. Last, we use a final softmax layer to obtain the predicted probability map for the segmentation results.

To make full use of the multi-modal knowledge and achieve better segmentation performance, we have tried different fusion strategies at different stages. In the developed IAFFNet, we have used the bi-directional Separation-and-Aggregation Gating operation (SAGate) fusion module between the two encoding branches to suppress noise, bidirectionally recalibrate the extracted features and fully exploit the complementary information. We propagate the average of the fused features and the modality-specific features to the next stage of the encoder for further feature transformation. At the decoding stage, we use adaptive fusion modules to effectively fuse the information transmitted from the encoder with the information in the decoder. At the decision level, we have developed an illumination classifier to help fuse the
results from the RGB skip-connections and the thermal skip-connections. The details of each component are described in the following sections.

A. BILATERAL FEATURE FUSION AT THE ENCODING STAGE

In previous works, features are encoded separately at the encoding stage [15], [17], [18], or fused features are only applied on one encoding branch [8], [9], [12]. Although attention modules are used in [12] and [13], they are applied on each encoding branch separately. To make full use of the multi-modal features at the encoding stage, we use an SAGate fusion module in the developed method. The complementary information from the two modalities are fused via the SAGate unit and then propagated along with the modality-specific features on each branch.

Separation-and-Aggregation Gating (SAGate) operation was previously used to suppress the noise of depth measurements and unify the RGB and depth information [20]. In this paper, we introduce SAGate to adaptively fuse the channel-wise and spatial-wise features from the RGB and thermal images. As shown in FIGURE 2, SAGate is a bi-directional symmetrical architecture that can fully use the fused features along both the RGB and thermal branches. First, the information from both the RGB and thermal images is concatenated and a global descriptor is obtained to extract expressive statistics by global average pooling operations.

A multilayer perceptron then learns the channel-wise attention weights, $W_{\text{RGB}}$ and $W_{T}$, are learned based on the concatenated recalibrated feature maps from both of the two branches. We obtain the spatial weights by two convolution layers and a sigmoid function. Finally, two spatially weighted feature maps are summed up to form the final fused feature map, $M_{\text{out}}$.

Suppose the two input feature maps are $\text{RGB}_{in}$ and $T_{in}$, the operations in the SAGate could be formulated as in (1)-(3). $\| \|$ denotes the concatenation of two feature maps. $F_{\text{gp}}$, $F_{\text{mlp}}$ and $F_{\text{conv} \times \times 1}$ refer to global average pooling, multilayer perceptrons and a $1 \times 1$ convolution layer respectively. $\sigma$ is a sigmoid function. Softmax indicates a softmax layer. $\otimes$ denotes a channel-wise multiplication, and $\odot$ is an element-wise multiplication.

$$T_{\text{rec}} = T_{in} + \text{RGB}_{in} \otimes \sigma(F_{\text{mlp}}F_{\text{gp}}(\text{RGB}_{in}||T_{in}))$$

$$\text{RGB}_{rec} = \text{RGB}_{in} + T_{in} \otimes \sigma(F_{\text{mlp}}F_{\text{gp}}(\text{RGB}_{in}||T_{in}))$$

$$W_{\text{RGB}}, W_{T} = \text{softmax}(F_{\text{conv} \times 1}(\text{RGB}_{rec}||T_{rec}))$$

$$M_{\text{out}} = \text{RGB}_{in} \odot W_{\text{RGB}} + T_{in} \odot W_{T}$$
B. ADAPTIVE FEATURE FUSION AT THE DECODING STAGE

Although a decoder can restore the original image size, some detailed features may be lost due to the downsampling operators. Skip-connecting the features from the encoding stage with the upsampled features at the decoding stage is usually used to achieve better segmentation performance. In previous works, the information from the encoding stage is propagated to the corresponding decoding layer, and simple concatenation [9], [13] or element-wise addition [17], [18] are usually used to fuse the features. To better select the channels with high discriminability, adaptive fusion modules are introduced to fuse the information from the encoder with the decoder in the developed method. Based on an advanced attention module [21], we have tested two adaptive fusion strategies in the evaluations.

A normalization-based attention module [21] is a computationally efficient but lightweight attention mechanism. To fuse the information from the encoder and decoder adaptively, we use a normalization-based channel attention sub-module in our network. As shown in FIGURE 3.

\[
X_{\text{trans}} = F_{\text{cbr}}(X_{\text{dec}}||X_{\text{skip}}) \quad (4) \\
X_{\text{out}} = X_{\text{trans}} \odot \sigma(w_{\gamma} \otimes F_{\text{BN}}(X_{\text{trans}})) \quad (5)
\]

in which \(w_{\gamma} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}\).

C. ILLUMINATION-AWARE FUSION OF DIFFERENT BRANCHES AT THE DECISION-LEVEL

Intuitively, RGB images contain more abundant information in the daytime, and thermal images provide more information about nighttime road scenes. Our experimental results also show that skip-connections from RGB or
thermal branches show different performance for day and night. Hence, we examine two explicit illumination-aware strategies to perform the fusion of different skip-connections in our evaluations.

To analyze the illumination conditions of each image, we have designed an illumination classifier. The classifier consists of two convolutional layers and three full-connection layers. A softmax layer is also used to get the final prediction results. Each convolutional layer is followed by a batch normalization layer, a rectified linear unit, and a max-pooling layer. Each full-connection layer is followed by a rectified linear unit and a dropout layer with a ratio of 0.5 to avoid overfitting. To reduce the computation load, the RGB images resized to 56 × 56 by bilinear interpolation are used as the input of the illumination classifier.

The outputs of the illumination classifier are $w_d \in (0, 1)$ and $w_n \in (0, 1)$, which indicate predicted probabilities that an image was captured in daytime or nighttime. As shown in FIGURE 4, in feature-level illumination-aware fusion, the features skipped from the RGB and thermal encoding branches are weighted based on the prediction results of the illumination classifier and then fused with the decoder features. In decision-level illumination-aware fusion, the segmentation results from different skip-connections are integrated based on the illumination classifier. If $w_d \geq w_n$, the image is predicted as captured during the daytime, and the results from the RGB skip-connections are used; otherwise, the image was captured in the nighttime, and the results from the thermal skip-connections are the final output.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL SETUP

1) DATASET

The dataset released by MFNet [6] has been widely used for the evaluations of RGB-T based semantic segmentation methods for autonomous driving in urban scenes. The dataset contains 820 pairs of daytime images and 749 pairs of nighttime scenes. The images were acquired using an InFReC R500 camera and the resolution is 640 × 480. Following the setting of [6], we used 784 pairs of images as the training set and 392 pairs for the validation set. We used the remaining images as the test set in our experiments. The training and validation sets both contain 1:1 images from the daytime and nighttime scenes.

2) IMPLEMENTATION DETAILS

Because it has shown good performance for RGB-T based semantic segmentation, we have used RTFNet [8] as the baseline in all our evaluations. We implemented the developed method based on the source code of [8]. All the evaluations were implemented using PyTorch 1.8.1 and CUDA 11.1, and the training were performed on a computer with an NVIDIA 3090 graphics card. For fair comparison, we follow the training settings in [8]. We used a stochastic gradient descent (SGD) solver for training. Momentum attenuation and weight attenuation are set to 0.9 and 0.0005, respectively. The initial learning rate was 0.01 for the segmentation network and 0.0001 for the illumination classifier. We used exponential decay schemes to gradually reduce the learning rate. The batch size is set as 2. We stopped network training until the loss remained stable.

3) EVALUATION METRIC

In the evaluations, $mIoU$ is used to evaluate the performance of the developed method quantitatively, which are the average intersection-over-union ($IoU$) values across all categories. The $IoU$ of each class is defined in (5). $TP$, $FP$, and $FN$ refer to the total number of true positives, false positives, and false negatives respectively. $i$ is the class index ranging from 1 to 9 including the background class.

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}$$

B. EVALUATIONS OF BILATERAL FEATURE FUSION AT THE ENCODING STAGE

In this experiment, we evaluated the SAGate module used at the encoding stage. The results are shown in TABLE 1. For fair comparison, we used RTFNet [8] as the baseline framework, and all methods in the experiment share the same decoding architecture. In the baseline method RTFNet [8], the features from RGB and thermal images are simply integrated by element-wise summation and propagated along one encoding branch as in [9] and [16]. As a recent example of the attention mechanism at the encoding stage, we also implemented the fusion module used in [13] and tested for comparison. In [13], RGB and thermal features are passed through a channel attention module separately and then added element-wisely. The fused features are propagated to the next encoding stage along the RGB branch.
As shown in TABLE 1, the developed method with SAGate fusion module has shown superior performance compared with other strategies. The overall performance in both the daytime and nighttime has improved. The mIoU of all-day has increased by 2.4% compared with the baseline method. While most other categories retain a similar performance, the categories of bike, car_stop and color_cone have improved greatly. These improvements benefit from the powerful feature integration ability of SAGate at the encoding stage.

Based on the concatenated features from both branches, adaptive fusion in the channel dimension can help suppress the noises from the input channels and highlight the discriminative channels for each image to form more representative features. With the adaptive spatial attention submodule, regions with significant clues can be emphasized and noisy regions may be suppressed. To fully use the recalibrated fused features, both the fused features and the modality-specific features are propagated to the next encoding stage along both branches.

C. EVALUATIONS ON THE ADAPTIVE FUSION AT THE DECODING STAGE

To validate the effectiveness of adaptive fusion at the decoding stage, several common fusion methods were carefully tested, and the results are listed in TABLE 2. For fair comparison, all the methods share the same encoding architecture of [8] with SAGate at the encoding stage. In the baseline method RTFNet+SAGate, no skip-connections are applied. As in [17] and [18], ⊕ uses an element-wise summation to fuse the skip-connections with the features at the decoding stage. ‘-RGB’ or ‘-T’ indicates whether the skip-connections are from the RGB or thermal branch of the encoding stage. Concatenation is also frequently used for the fusion of the skip-connections with the decoding branch [9], [13]. The ε +BN-RGB and ε +BN-T variants represent concatenation followed by a batch normalization layer. Two developed adaptive fusion strategies are weight-then-skip and skip-then-weight.

As we can see, both the source of the skip-connections and the fusion methods may affect the skip-connections. Element-wise summation and concatenation have shown similar performance. Skip-connections from the RGB branch have shown an overall improvement in both cases, and skip-connections from the thermal branch have degraded the segmentation performance. In weight-then-skip, only the thermal branch has shown a slight improvement. skip-then-weight has shown the best feature integration ability. With either RGB branch or the thermal branch, skip-connections have shown positive effects on segmentation performance. Compared with the baseline, the segmentation performance under
TABLE 3. Comparison of different illumination-aware strategies (mIoU).

| Fusion methods of different skip-connections | car | person | bike | curve | car_stop | guardrail | color_cone | bump | Day | night | All-day |
|---------------------------------------------|-----|--------|------|-------|---------|-----------|------------|------|-----|-------|---------|
| skip-then-weight-RGB                        | 84.8| 72.4   | 62.1 | 47.0  | 26.7    | 6.7       | 45.3       | 61.9 | 50.2| 52.7  | 56.1    |
| skip-then-weight-T                          | 86.1| 72.3   | 63.9 | 47.0  | 24.5    | 3.9       | 51.5       | 58.3 | 48.8| 54.9  | 56.2    |
| Feature-level illumination-aware fusion     | 86.4| 72.3   | 60.9 | 43.7  | 26.6    | 1.2       | 32.4       | 50.6 | 46.3| 51.1  | 52.4    |
| Decision-level illumination-aware fusion    | 86.7| 72.6   | 64.3 | 47.6  | 27.3    | 3.9       | 47.7       | 61.4 | 50.2| 54.9  | 56.6    |

different illuminations have also shown consistent improvement. The results show that, the features from different levels have been more effectively fused in skip-then-weight. The channels of features with high discriminability from different levels have been adaptively selected to guide the segmentation of each image.

From the skip-then-weight fusion results in TABLE 2, we also note that the connections from the RGB branch and the thermal branch show quite different performance under different illumination scenarios. As we can see, skip-connections from the RGB branch tend to obtain better performance in the daytime, and skip-connections from the thermal branch have shown more benefits in the nighttime. The reason may be straightforward. With sufficient illumination, we can usually obtain more abundant information about the road scenarios from the RGB camera in the daytime. In the nighttime, the thermal cameras show superiority on objects with radiation heat. Based on this observation, we have performed a careful study on the fusion of different skip-connection branches in the next section.

D. EVALUATIONS ON THE ILLUMINATION-AWARE STRATEGIES

Based on the observations in Section IV-C, skip-connections from the RGB and thermal branch show different advantages in the daytime and nighttime. In this experiment, we have studied two illumination-aware strategies to fuse different skip-connections, and the results are shown in TABLE 3.

As we can see, when appropriate fusion strategies are adopted, illumination knowledge can further boost the semantic segmentation performance. The overall mIoU of all-day has been increased compared with either skip-connections by the decision-level illumination-aware fusion method. Among the two illumination-aware fusion strategies, the results show that decision-level illumination-aware fusion can exploit the complementary relationships of different skip-connections and perform much better than the feature-level fusion methods. The reason may be that high-level multi-modal fusion is more suitable for the high-level semantic segmentation tasks.

E. COMPARISON WITH OTHERS

To validate the efficiency of the developed method, we also performed a comparison with other methods based on RGB and thermal images. We compared the proposed IAFFNet with RTFNet [8], FuseSeg [9], MFNet [6], UNet [22], SegNet [1], FuseNet [23], PSTNet [10], ABMDRNet [15], MLFNetR34 [13], FEANet [12], and EDGE-Aware [18]. We obtained the results for RTFNet [8] from our tests based on the source code provided by the authors. We obtained the results for UNet [22], SegNet [1], FuseNet [23], and MFNet [6] from [8], and others from their own publications. As shown in TABLE 4, through carefully designed fusion strategies at different stages, mIoU has increased by 5.38% in the developed method compared with the baseline RTFNet, which validates the efficiency of the developed strategies. Almost every category has shown significant improvement. Compared with other recent networks, our method has also shown its superiority in terms of mIoU. The person category and the bump category have achieved the best results.

To show the performance of different methods under different illumination conditions, TABLE 5 also lists the results of some methods in the daytime and nighttime respectively. Those methods which have not provided the results under different illumination conditions in their publications are not included in TABLE 5. We observe that, with the careful design of fusion strategies at different stages, the segmentation performance under different illumination conditions has improved compared with the baseline method [8]. The mIoU in both daytime and nighttime has achieved the state-of-the-art. With the bidirectional channel-wise and spatial-wise attention mechanism, the information from RGB and thermal images have been effectively fused at the encoding stage. By adaptively fusing the detailed features from the skip-connections with the high-level features in the decoder, the method can better use features from different levels and...
TABLE 4. Comparison with other methods throughout all the day (mIoU).

| Methods                  | car | person | Bike | curve | car_stop | guardrail | color_cone | bump | All-day |
|--------------------------|-----|--------|------|-------|----------|-----------|------------|------|---------|
| Unet(4c)[22]             | 66.2| 60.5   | 46.2 | 41.6  | 17.9     | 1.8       | 30.6       | 44.2 | 45.1    |
| Unet(3c)[22]             | 65.2| 42.6   | 47.8 | 27.8  | 20.8     | 0.0       | 35.8       | 31.0 | 40.8    |
| SegNet(4c)[1]            | 65.3| 55.7   | 51.1 | 38.4  | 10.0     | 0.0       | 12.0       | 51.5 | 42.3    |
| SegNet(3c)[1]            | 57.3| 27.1   | 49.9 | 16.8  | 0.0      | 0.0       | 0.0        | 37.7 | 31.7    |
|FuseNet[23]               | 75.6| 66.3   | 51.9 | 37.8  | 15.0     | 0.0       | 21.4       | 45.0 | 45.6    |
| MFNet[6]                 | 65.9| 58.9   | 42.9 | 29.9  | 9.9      | 0.0       | 25.2       | 27.7 | 39.7    |
|FuseSeg[9]                | **87.9**| 71.7   | 64.6 | 44.8  | 22.7     | 6.4       | 46.9       | 47.9 | 54.5    |
|PSTNet[10]                | 76.8| 52.6   | 55.3 | 29.6  | 25.1     | 15.1      | 39.4       | 45.0 | 48.4    |
|ABMIDRnet[15]            | 84.8| 69.6   | 60.3 | 45.1  | 33.1     | 5.1       | 47.4       | 50.0 | 54.8    |
|MILNet[34][13]           | 82.3| 68.1   | **67.3**| 27.3  | 30.4     | **15.7**  | **55.6**   | 40.1 | 53.8    |
|FEANet[12]               | 87.8| 71.1   | 61.1 | 46.5  | 22.1     | 6.6       | 55.3       | 48.9 | 55.3    |
|EDGE-Aware[18]           | 87.6| 69.8   | 58.8 | 42.8  | **33.8** | 7.0       | 48.3       | 47.1 | 54.8    |
|RTFNet(baseline)[8]      | 86.9| 69.0   | 60.3 | 47.0  | 17.2     | 0.5       | 28.4       | 55.5 | 51.4    |
|IAFFNet(ours)            | **86.6**| **72.6**| 64.3 | **47.6**| 27.3     | 3.9       | 47.7       | **61.4**| **56.6**|

FIGURE 5. Qualitative segmentation results. Columns 1-4 are daytime scenarios, and columns 5-8 are nighttime scenarios.

obtain more accurate segmentation results. Finally, with the decision-level illumination-aware strategies, the complementary advantages of the RGB and thermal images are further exploited.

FIGURE 5 displays some qualitative results. The first four columns are from the daytime scenes, and the last four columns are from the nighttime scenes. Compared with the baseline method RTFNet [8], the developed network has significantly reduced miss detections and misclassifications. In the first column, a number of guardrail pixels have been misclassified in [8] and the developed method has obtained better results. In the sixth column, a faraway vehicle and a small color cone have been correctly identified in the developed method. More accurate boundaries have also been obtained, such as the people in the second, fourth, and last columns. These may benefit from the developed adaptive skip-connection module to help with the fusion of both semantic and detailed features. In the seventh column, illuminations in the image are uneven. The person at the left shows more clues in the thermal image, and the bicycles are more obvious in the RGB image. In the developed method, persons and bicycles are both well detected, which shows the powerful ability of the developed method to fuse multi-modal information.
In addition, in the second and fifth column, curves have been correctly found even though they are not manually annotated in the ground truth. Hence, higher mIoU may be expected if corrections are made in the ground truth.

V. CONCLUSION

Through a careful study of the fusion strategies at different stages, we have developed an RGB-T semantic segmentation framework called IAFFNet. We have introduced SAGate into the encoder to exploit the multi-modal features fully at the encoding stage. We have adaptively fused the information from the encoder and decoder, and we have used a decision-level illumination-aware fusion has to further improve the robustness of the developed method in daytime and nighttime.

In the future, we will improve the efficiency of the developed method to achieve real-time application. Multi-label supervision [16], [17], [18] can also be considered to help improve the boundary accuracy of segmentation.

REFERENCES

[1] 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, Jan. 2017.

[2] 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.

[3] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” 2017, arXiv:1704.06857.

[4] 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.

[5] 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.

[6] 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.

[7] M. O. Aldana and M. H. R. Diaz, “Infrared thermal imaging fundamentals, research and applications,” Latin-Americ. J. Phys. Educ., vol. 5, no. 1, p. 42, 2011.

[8] 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.

[9] 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. 99, pp. 1–12, Jul. 2020.

[10] 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), May 2020, pp. 9441–9447.

[11] X. Lan, X. Gu, and X. Gu, “MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation,” Int. J. Speech Technol., vol. 52, no. 5, pp. 5817–5829, Mar. 2022.

[12] F. Deng, “FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2021, pp. 4467–4473.

[13] Z. Guo, X. Li, Q. Xu, and Z. Sun, “Robust semantic segmentation based on RGB-thermal in variable lighting scenes,” Measurement, vol. 186, Dec. 2021, Art. no. 110176.

[14] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[15] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2633–2642.

[16] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, “MFFNet: Multi-scale feature fusion and enhancement network for RGB-thermal urban road scene parsing,” IEEE Trans. Multimedia, vol. 24, pp. 2526–2538, 2022.

[17] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, “GMNet: Graded-feature multi-label-learning network for RGB-thermal urban scene semantic segmentation,” IEEE Trans. Image Process., vol. 30, pp. 7790–7802, 2021, doi: 10.1109/TIP.2021.3109518.

[18] W. Zhou, S. Dong, C. Xu, and Y. Qian, “Edge-aware guidance fusion network for RGB-thermal scene parsing,” in Proc. AAAI Conf. Artif. Intell., 2022, pp. 3571–3579.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[20] X. Chen, “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 561–577.

[21] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, “NAM: Normalization-based attention module,” 2021, arXiv:2111.12419.

[22] 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.

[23] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture,” in Proc. Asian Conf. Comput. Vis. Berlin, Germany: Springer, 2016, pp. 213–228.

TABLE 5. Results of the recent methods on daytime and nighttime images (mIoU).

| Methods              | Daytime | Nighttime | All-day |
|----------------------|---------|-----------|---------|
| Unet(4c) [22]        | 38.0    | 44.0      | 45.1    |
| Unet(3c) [22]        | 37.5    | 37.0      | 40.8    |
| SegNet(4c) [1]       | 34.6    | 41.7      | 42.3    |
| SegNet(3c) [1]       | 29.5    | 27.4      | 31.7    |
| FuseNet[23]          | 41.0    | 43.9      | 45.6    |
| MFNet[6]             | 36.1    | 36.8      | 39.7    |
| FuseSeg[9]           | 47.8    | 54.6      | 54.5    |
| MLFNetR34[13]        | 45.6    | 55.0      | 53.8    |
| RTFNNet(baseline) [8] | 45.6   | 52.2      | 51.4    |
| IAFFNet(ours)        | 50.2    | 54.9      | 56.6    |
Y.-L. Hou et al.: IAFFNet for All-Day RGB-Thermal Semantic Segmentation of Road Scenes

YA-LI HOU (Member, IEEE) received the B.Eng. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2004, the master’s degree from the Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China, in 2007, and the Ph.D. degree from The University of Hong Kong, Hong Kong, SAR, China, in 2011. She is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. Her research interests include pattern recognition and computer vision, with a special focus on object detection, semantic segmentation, and instance segmentation.

YAN JIA was born in Heze, Shandong, China, in 1997. She received the B.S. degree in electrical and intelligent building from the Qingdao University of Technology, Qingdao, China, in 2019. She is currently pursuing the M.S. degree with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China.

ZHIJIANG HOU received the master’s degree in pattern recognition and intelligent systems from Nankai University, Tianjin, China, in 2009. He is currently with the Library, Tianjin University of Technology, Tianjin. His main research interests include intelligent text processing and literature mining.

XIAOLI HAO received the B.E., M.E., and Ph.D. degrees from the School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China, in 1992, 1995, and 2010, respectively. She joined the School of Electronics and Information Engineering, Beijing Jiaotong University, in 1995, and became an Associate Professor, in 2002. In 2006, she was a Visiting Scholar with the University of California at San Diego, San Diego, CA, USA. Her current research interests include optical imaging, signal processing, and machine vision.

YAN SHEN (Member, IEEE) was born in Xi’an, Shaanxi, China, in 1979. She received the B.S. and M.S. degrees in telecommunication engineering from Xidian University, Xi’an, in 2001 and 2004, respectively, and the Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2007. She is currently an Associate Professor and the Tutor of Ph.D. student with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing. She is the author of three books, more than 40 articles, and more than ten inventions and software copyrights. Her research interests include deep learning, image restoration, sparse representation, and object detection.

* * *

VOLUME 10, 2022

129711