Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion

Jiawei Li
University of Science and Technology Beijing
Beijing, China
ljw19970218@163.com

Jiansheng Chen[*]
University of Science and Technology Beijing
Beijing, China
jschen@ustb.edu.cn

Jinyuan Liu
Dalian University of Technology
Dalian, China
atlantis918@hotmail.com

Huimin Ma
University of Science and Technology Beijing
Beijing, China
mhmpub@ustb.edu.cn

Fig. 1. Fusion, detection and segmentation comparisons with state-of-the-art methods on M³FD and MFNet datasets. We can obviously notice the superiority of our IGNet in the zoomed-in patches and radar plots.

ABSTRACT

Infrared and visible image fusion has gradually proved to be a vital fork in the field of multi-modality imaging technologies. In recent developments, researchers not only focus on the quality of fused images but also evaluate their performance in downstream tasks. Nevertheless, the majority of methods seldom put their eyes on mutual learning from different modalities, resulting in fused images lacking significant details and textures. To overcome this issue, we propose an interactive graph neural network (GNN)-based architecture between cross modality for fusion, called IGNet. Specifically, we first apply a multi-scale extractor to achieve shallow features, which are employed as the necessary input to build graph structures. Then, the graph interaction module can construct the extracted intermediate features of the infrared/visible branch into graph structures. Meanwhile, the graph structures of two branches interact for cross-modality and semantic learning, so that fused images can maintain the important feature expressions and enhance the performance of downstream tasks. Besides, the proposed leader nodes can improve information propagation in the same modality. Finally, we merge all graph features to get the fusion result. Extensive experiments on different datasets (*i.e.*, TNO, MFNet, and M³FD) demonstrate that our IGNet can generate visually appealing fused images while scoring averagely 2.59% mAP@.5 and 7.77% mIoU higher in detection and segmentation than the compared

state-of-the-art methods. The source code of the proposed IGNet can be available at https://github.com/lok-18/IGNet.

KEYWORDS

Image fusion, graph neural network, cross-modality interaction, leader node

1 INTRODUCTION

Due to the inadequacy of single-modality imaging, the resulting images are commonly defective in complex scenes [22], [21]. As a representative, Visible images are more in line with the human visual system (HVS), but susceptible to environmental factors. In this case, researchers attempt to fuse visible images with ones of another modality to counteract the disadvantages of single-modality imaging. Complementarily, infrared images can capture salient targets with thermal radiation sensors. Texture details and resolution of them often perform undesirably. Therefore, infrared and visible image fusion (IVIF) emerges as the times require, which can possess information from different modalities simultaneously. Acting as an indispensable part of multi-modality imaging technology, IVIF has drawn extensive attention to computer vision tasks, *e.g.*, vehicle detection [32], video surveillance [27] and image stitching [8].

For the past decade, deep learning networks have been introduced to explore the IVIF task [43], which mainly contains convolution neural network (CNN)-based [19] and transformer-based methods [25]. These methods focus on accurate feature extraction

for inputs while promoting the fusion efficiency significantly. Compared with previous traditional approaches, deep learning-based methods can utilize more efficient feature extraction capabilities to obtain fusion results with higher efficiency. With further development, researchers also pay attention to the performance of down-stream tasks after fusion [33]. That is to say the results of down-stream tasks are closely related to the fusion images.

Existing mainstream IVIF methods have reached a certain height, nevertheless, there are still several drawbacks: (i) the uneven distribution of infrared and visible information extracted from networks causes fusion results only to be biased towards one modality [14], which can not perform the prominent regions of source images well. (ii) since feature learning often acts separately on each single branch, the information contained in networks may lack the communication of cross modalities [34]. (iii) the internal design of message delivery is not well taken into account in several networks [44], so some significant details of source images can not be displayed in fused results.

To alleviate the drawbacks mentioned above, in this paper, we propose an interactive GNN-based architecture between cross modality for the IVIF task, termed as IGNet. Concretely, multi-scale shallow features are first extracted by convolutions and the proposed structure salience module (SSM). Then, we construct a graph interaction module (GIM) to obtain graph structures of different branches for feature learning. Note that the interaction of cross-modality graph features enables the proposed IGNet to achieve more semantic information, which can improve the performance of down-stream tasks, *e.g.*, object detection, and image segmentation. In addition, the establishment of leader nodes guides the message propagation effectually to avoid image quality degradation caused by feature loss. Fig. 1 proves that our proposed IGNet maintains the superior position regardless of subjective visual results or objective marks compared with state-of-the-art methods.

In brief, the contributions can be divided into the following aspects:

- For optimizing the internal relationship of fusion and down-stream (*i.e.*, object detection and image segmentation) tasks, to the best of our knowledge, we are the first to apply GNN into the IVIF method. To this end, the fused results can contain faithful visual representation and feature comprehension abilities.

- We propose a graph interaction module (GIM) for getting graph structures. It can proceed cross-modality communication through graph features, which highlight the desired details of fusion results. Furthermore, the semantic-wise information can also be extracted by GIM for improving down-stream results.

- Unlike the common GNN, the leader nodes are employed for information delivery after achieving graphs. Accompanied by a leader node as a pioneer, fusion images can maintain abundant textures from source inputs.

- We conduct image fusion, detection, and segmentation experiments on TNO, M^{3}FD, and MFNet datasets. Compared with the other seven state-of-the-art approaches, our proposed IGNet performs foremost in all tasks.

2 RELATED WORKS

2.1 Infrared and Visible Image Fusion

Deep learning has promoted rapid development in the field of image fusion [12], [20], [9], [13]. In early stages, researchers are dedicated to improving the performance of fused images by CNN-based methods, which are mainly divided into three classes, *i.e.*, End-to-End-based models [19], [23], Encoder-Decoder models [10], and generative adversarial network (GAN)-based models [26].

More specifically, End-to-End models preset parameters before unsupervised training [14]. Liu *et al*. [19] proposed a coarse-to-fine deep network with an end-to-end manner to learn multi-scale features from infrared and visible images. The structure details were also refined by the proposed edge-guided attention mechanism. The Encoder-Decoder models need to design a fusion rule to integrate features extracted from the encoder, and then output the fusion results from the decoder [39]. Zhao *et al*. [47] conducted a novel encoder to decompose source images into background and detail feature maps, which can highlight targets, especially in the dark. The GAN-based models require a generator and a discriminator for adversarial learning. Li *et al*. [11] effectively combined the attention mechanism with GAN, namely AttentionFGAN. Moreover, extensive transformer-based models have also received much attention in the IVIF task [25]. Tang *et al*. [25] utilized Swin Transformer and cross-domain long-range learning into the IVIF task, which connected local features with global representation.

To further explore the performance of fusion images, researchers have introduced down-stream tasks *e.g.*, object detection and image segmentation, into the IVIF task. As a representative, Liu *et al*. [18] proposed a unified architecture and built a multi-modality dataset for image fusion and detection. Sun *et al*. [31] employed the information back-propagated by detection loss in the proposed network to obtain fused images with excellent detection results. For getting more semantic features, Tang *et al*. [34] proposed a cascaded structure called SeAFusion, which connects the fusion network with a pretrained segmentation module. Zhao *et al*. [46] conducted a novel two-stage training mode for fusion. The detection and segmentation results also performed well in this benchmark.

2.2 Graph Neural Network

In recent years, GNN-based approaches have become increasingly popular in computer vision. Different from traditional CNN-based methods, the unique structure of GNN enables to extract and transfer more efficient features. Therefore, GNN is commonly implemented in the feature-wise tasks. As a representative, Xie *et al*. [38] proposed a scale-aware network with GNN to conduct few-shot semantic segmentation. In the medical field, Huang *et al*. [6] employed a semi-supervised network for medical image segmentation, which could help doctors diagnose diseases better. Recently, GNN becomes popular in the field of saliency detection. It can effectively highlight the salient mask of measured targets. Specifically, Luo *et al*. [24] tried to cascade graph structures for salient object detection (SOD) with RGB-D images. Song *et al*. [30] devised a multiple graph module to realize the RGB-T SOD task. GNN can be also applied in Co-Saliency Detection (CSD) and Instance Co-Segmentation (ICS). Li *et al*. [15] presented a general adaptive GNN-based module to deal with CSD and ICS. In addition, some low-level tasks can also

Fig. 2. Pipeline of the proposed IGNet. Specifically, we feed multi-scale features into the graph interaction module (GIM) for generating graph structures in different modalities. The cross-modality interaction between graphs is depicted in detail. The leader nodes guide the information delivery from one graph to the latter. Note that we construct graphs in the infrared/visible branch with three loops, respectively. The bottom row represents the legend of the component.

perform well by using GNN as their benchmark. Li *et al.* [16] proposed a novel GNN-based method for image denoising. In summary, GNN maintains sensitivity to semantic information, while handling pixel-level tasks well. Hence, our proposed IGNet can exploit the advantages of GNNs for deeper exploration of IVIF tasks, which can simultaneously improve the quality of fused images and the performance of corresponding downstream tasks.

3 METHOD

3.1 Motivation

In the IVIF task, networks often extract features in infrared and visible branches separately, while ignoring the interaction between modalities. It may cause textures of source images can not be completely displayed in fusion results. With the information delivery during training, the occurrence of feature forgetting is inevitable as well. Besides, the fused images will directly affect the performance of the down-stream results. There is no doubt that applying an effective architecture to achieve visually appealing images can improve the accuracy of detection and segmentation. Significantly, how to obtain fused images with prominent targets, fine textures, and rich semantic information is the key to handling the above issues. Hence, it is our motivation to realize a general IVIF framework, which can obtain fusion and semantic information in pixel and feature domains concurrently.

3.2 Overall Workflow

The proposed IGNet adopts a dual-branch framework in the feature learning stages. Subsequently, we aggregate the infrared and visible branches to achieve fusion images. The overall pipeline is illustrated in Fig. 2. To be specific, two different-scale features (*i.e.*, f_1^* and f_2^*) can be generated by the first two convolutional layers, where $*$ denotes the infrared/visible branch. Then, we modify f_2^* through the SSM for getting the salient-structure feature f_3^*. It is formulated as follow:

$$f_3^* = \mathcal{S}(f_2^*), \tag{1}$$

where $\mathcal{S}$ means the SSM. For constructing connections between source images, f_i^* is fed into the GIM to build a learnable graph structure with three loops. We define this process as follows:

$$g_* = \sum_{i=1}^{3} \mathcal{G}(f_i^*), \tag{2}$$

where $i \in \{1, 2, 3\}$, g_* denotes graph features and $\mathcal{G}$ is the GIM. At last, we combine decorated features to achieve final fusion results:

$$I_f = Conv\big(Concat(g_{ir}, g_{vis})\big), \tag{3}$$

where I_f means fused images. $Conv(\cdot)$ and $Concat(\cdot)$ represent convolution and concatenate operations, respectively. Moreover, the employed loss function can effectively transfer information through back-propagation, which is also explicated in Section. 3.5.

3.3 Structure Salience Module

As shown in Fig. 2, we use the SSM to optimize f_2^*, deepening the expression of deep structure features. After passing through a convolutional layer, the SSM conducts Maxpooling and Avgpooling to coordinate detailed patches and global information simultaneously. We use Element-wise Multiplication to deal with the two pooling information, which can excavate more salient contents from infrared images. Since more detailed textures are contained in the visible branch, Element-wise Addition is exploited to enrich the overall perception instead.

Inspired by SENet [5], we also introduce attention to the SSM. Firstly, the aforementioned feature is flattened by Global Average Pooling (GAP). Secondly, we assign two Fully Connected Layers and Sigmoid to generate the corresponding channel weight. It can not only increase the salience of feature representation but also highlight parts that conform to HVS in fused images. Finally, we multiply the feature with channel weight to achieve salient-structure feature f_3^*.

(a) Node generation (b) Edge generation

Fig. 3. Specific illustration of (a) node generation and (b) edge generation.

3.4 Graph Interaction Module

We design a graph structure for information learning and interaction between different modalities in GIM, which can improve the quality of fusion results. Furthermore, it enables images to contain more high-level information, so that the down-stream tasks (*i.e.*, object detection, image segmentation) also perform well. The middle of Fig. 2 shows the specific workflow of the GIM.

As the infrared branch an example, we provide the former features f_i^{ir} with different scales acting as pioneer factors to the GIM for graph generation. Note that the GIM implements three loops of graph structures with three nodes in each branch to balance the performance of fusion results and operational efficiency. Detailed ablation experiments are conducted in Section. 4.6. In the process of creating graphs, we connect nodes of different scales from the same modality and nodes of the same scale from different modalities concurrently. The interactive way can restrict information imbalance while enhancing the representation of each input in fused images. After obtaining a graph, nodes constitute a corresponding leader node g_i^{ir} to guide information delivery for the latter graph. Owing to the assistance of leader nodes, the GIM can resist information loss, improving the capability of feature learning. The leader nodes g_1^{ir}, g_2^{ir} and g_3^{ir} are finally mixed together to achieve the graph feature g_{ir}.

3.4.1 Node and Edge Generation. Aimed at ensuring the diversity of features, we divide them into nodes of different scales through the pyramid pooling module (PPM) [45]. Fig. 3 (a) describes the detailed process of node generation. We employ pyramid pooling, convolution, and upsample operations to split f_i^* into multiple scales to obtain the nodes in the graph, respectively. Note that the nodes and f_i^* keep consistent except for the number of channels. This process can be proved as follow:

$$(g_i^*)_o = Up\Big(Conv\big(\mathcal{P}(f_i^*)\big)\Big), \qquad (4)$$

where $(g_i^*)_o$ represents the o-th ($j \in \{1, 2, 3\}$) node of the i-th ($i \in \{1, 2, 3\}$) graph in * (infrared/visible) modality. Up and $\mathcal{P}$ denote the upsample and pyramid pooling operations.

The production of edges in Fig. 3 (b) also stands an essential role of the graph generation, which carries the information transmission between nodes. We build edges in different-scale nodes from the same modality. Distinctively, nodes with the same scale from different modalities are linked for learning more semantic-level relations. The edge generation in g_j and g_k is bidirectional and defined as:

$$e_{j,k} = Conv(g_j - g_k), \qquad (5)$$

(a) Leader node generation (b) Information delivery

Fig. 4. Specific illustration of (a) leader node generation and (b) information delivery.

$$e_{k,j} = Conv\big(\mathcal{N}(g_j - g_k)\big), \qquad (6)$$

where $\mathcal{N}$ means the negation operation. $e_{j,k}$ ($e_{k,j}$) is the edge embedded from g_j (g_k) to g_k (g_j). In addition, the message passing $m_{j,k}$ can be formulated as:

$$m_{j,k} = Sigmoid(e_{j,k}) \cdot g_j, \qquad (7)$$

where *Sigmoid* denotes the Sigmoid operation.

3.4.2 Leader Node and Information Delivery. In Fig. 4 (a), the introduction of leader nodes makes the delivery of semantic information between nodes in the graph more effectively, which can be represented as follow:

$$g_i^* = Conv\Big(Concat\big((g_i^*)_1, (g_i^*)_2, (g_i^*)_3\big)\Big) \qquad (8)$$

In the process of information delivery as shown in Fig. 4 (b), the leader node generates the corresponding leader weight by the GAP and Sigmoid operation. After three former nodes pass through the convolutions, we multiply them with the leader weight in channel domain. Finally, the extracted multi-level features are propagated into the latter nodes, which can embody both details and targets clearly in fused images.

3.5 Loss Function

To guarantee that more meaningful information can be learned during the training phase, we introduce three varieties of loss functions, *i.e.*, the pixel loss $\mathcal{L}_{MSE}$, the edge loss $\mathcal{L}_{edge}$ and the structure loss $\mathcal{L}_{SSIM}$. The combined $\mathcal{L}_{total}$ can be shown as follow:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + \alpha \mathcal{L}_{edge} + \beta \mathcal{L}_{SSIM}, \qquad (9)$$

where α and β are preset hyperparameters with the value of 10 and 0.5. Specifically, mean squared error (MSE) can measure the pixel intensity between source images and the fusion result. Note that we conduct weighted average to source images before calculating. It can be defined as:

$$\mathcal{L}_{MSE} = MSE\big((I_{ir} + I_{vis})/2, I_f\big), \qquad (10)$$

where I_{ir} and I_{vis} mean infrared and visible images, respectively. To highlight edge details, $\mathcal{L}_{edge}$ selects the infrared/visible image with a larger gradient value to achieve the edge gradient:

$$\mathcal{L}_{edge} = \| \nabla I_f - max(\nabla I_{ir}, \nabla I_{vis}) \|_1^2, \qquad (11)$$

where ∇ is the gradient operator and $\| \cdot \|_1$ is the l_1-norm. Besides, structural similarity index measure (SSIM) [37] can calculate the similarity between source images and the fusion image, which is expressed as follow:

$$\mathcal{L}_{SSIM} = \big(1 - SSIM(I_f, I_{ir})\big) + \big(1 - SSIM(I_f, I_{vis})\big). \qquad (12)$$

With the help of the above loss function, the pixel and structural level information can be fully retained, which makes the fusion and down-stream results perform well.

4 EXPERIMENTS

In this section, we first introduce the experimental setup, comparison approaches and dataset selection. Then, we analyze the fusion, detection, and segmentation results separately to verify the superiority of our proposed method. Furthermore, ablation experiments are mentioned to demonstrate the effectiveness of the proposed modules.

4.1 Experimental Implementation

In the training phase, we choose Adam optimizer to adjust the training parameters, where the stride and bitch size are set to 8 and 2. The initial learning rate of the network is $1e^{-3}$ with a decay rate of $2e^{-4}$. The total epoch is 100. In the loss function, the hyperparameters α and β are set to 10 and 0.5, respectively. The selection of training datasets is presented in Section. 4.2. All experiments are implemented on an NVIDIA GeForce 3070Ti GPU with PyTorch framework.

4.2 Dataset Selection and Comparison Approaches

The TNO [35], M^3FD [18] and MFNet [3] datasets contain plenty of infrared and visible image pairs. Moreover, the M^3FD and MFNet datasets also have image pairs that have been labeled for detection and segmentation. Before training, we combine 15 TNO pairs, 150 M^3FD pairs and 1083 MFNet pairs as the training set of our IGNet. The testing set consists of 10 TNO pairs, 150 M^3FD pairs, and 361 MFNet pairs. Note that the division of the TNO and M^3FD datasets is random, the MFNet dataset is based on [34].

We select seven state-of-the-art methods including DIDFuse [47], U2Fusion [40], SDNet [42], TarDAL [18], UMFusion [36], DeFusion [17] and ReCoNet [7], to compare with our proposed IGNet in qualitative and quantitative results. During the fusion task, we apply six evaluation metrics, *i.e.*, entropy (EN), visual information fidelity (VIF) [4], average gradient (AG), correlation coefficient (CC) [29], the sum of the correlations of differences (SCD) [1] and edge-based similarity measure ($Q_{ab/f}$) [41], for objective estimation. Larger values of the above-mentioned metrics mean the image quality performs better.

In the detection task, 4200 pairs of labeled images are employed as training, validation, and testing sets in a ratio of 8:1:1. The labels are marked into six categories, *i.e.*, people, bus, car, motorcycle, truck, and lamp. A mainstream detector, YOLOv5 [28], is conducted for detection. We set the optimizer, learning rate, epoch, and batch size as SGD optimizer, $1e^{-2}$, 400, and 8, respectively. The mAP@.5 is measured for quantitative comparison. Moreover, we utilize DeepLabV3+ [2] to segment fusion results, which choose the MFNet dataset as training and testing sets. There are nine labels in the sets, including background, car, person, bike, curve, car stop, guardrail, color cone, and bump. The training epoch and bitch size are set as 300 and 8, while other parameters keep the same as in the original experiment. The mIoU is selected for objective evaluation. In summary, we realize fusion images of each comparison approach

to retain down-stream tasks, then analyze their corresponding performance.

4.3 Analysis for Fusion Results

4.3.1 Qualitative Analysis. We depict qualitative results on TNO, MFNet, and M^3FD datasets in Fig. 5. Obviously, our results outperform other state-of-the-art methods. For instance, targets and surrounding scenes obscured by smoke can be clearly displayed on the TNO dataset. In the second illustration, TarDAL and ReCoNet occur over-exposed regions, while U2Fusion, SDNet and UMFusion remain low-contrast performance. Although DIDFusion can highlight luminance information (*e.g.*, car lights), its background abandons many texture details, which is unfriendly to HVS. In addition, benefiting from the cooperation of GNN, blur artifacts can be effectively mitigated as shown in the green enlarged patch of the third row.

4.3.2 Quantitative Analysis. In Table. 1, we enumerate the mean scores for the six metrics in the three testing sets. From an overall perspective, the quantitative results of our method stand in the lead position. Specifically, CC and SCD achieve the highest scores, which indicates the mutual connection between our fusion images and source inputs is the tightest. The highest value of $Q_{ab/f}$ reflects that the edge contours of targets can be well represented. Moreover, the higher performance of EN and AG demonstrates that a large amount of information is preserved in our fusion results. Since our approach pays greater emphasis on information delivery, the VIF value also keeps at a higher level.

4.4 Analysis for Detection Results

4.4.1 Qualitative Analysis. As shown in Fig. 6, the disturbance of environmental factors causes the detection results of single-modal images to be generally weaker than those of fusion results. However, the sensitivity of different fusion results to detection is also varied. In the first row of examples, SDNet and DeFusion present significantly low confidence and error detection regions, which may mislead observers. Moreover, "Truck" is detected as "Car" in the second set, while missing detection of cars in the corner also occurs. As a representative, our fusion results contain rich advanced features, so that the corresponding detection results can avoid the above phenomena. We can also notice that our detection results achieve high-confidence scores on all labeled categories.

4.4.2 Quantitative Analysis. Table. 2 exhibits the AP@.5 of each label and matching total mAP@.5 measured by detection results of fused images, which can obtain higher indicators than single infrared or visible images. Under the comparison of fusion methods, our proposed IGNet performs 2.59% higher than others in detection. It is worth noting that IGNet can not only achieve excellent detection results but also take into account the quality of fusion images.

4.5 Analysis for Segmentation Results

4.5.1 Qualitative Analysis. Visual results of the segmentation on the MFNet dataset are presented in Fig. 7. Similarly, we also employ fusion results of each method as the input to obtain segmentation results. Due to less semantic information contained in images,

Infrared image   Visible image   DIDFuse   U2Fusion   SDNet   TarDAL   UMFusion   DeFusion   ReCoNet   IGNet

Fig. 5. Visual comparisons of different approaches on TNO, MFNet and M³FD datasets. Our proposed IGNet can achieve notable targets and fine background details. The enlarged red and green circles are detailed patches of fusion results.

Table 1

QUANTITATIVE COMPARISONS OF OUR IGNET WITH SEVEN STATE-OF-THE-ART METHODS ON TNO, MFNET AND M³FD DATASETS. OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED IN RED AND BLUE, RESPECTIVELY.

Method	Dataset:TNO						Dataset:MFNet						Dataset:M³FD					
	EN	VIF	AG	CC	SCD	$Q_{ab/f}$	EN	VIF	AG	CC	SCD	$Q_{ab/f}$	EN	VIF	AG	CC	SCD	$Q_{ab/f}$
DIDFuse	7.066	0.738	5.150	0.503	1.726	0.413	2.695	0.277	2.005	0.526	1.007	0.176	7.108	0.879	5.663	0.558	1.666	0.482
U2Fusion	6.844	0.663	5.062	0.242	1.739	0.444	4.612	0.503	2.899	0.627	1.262	0.364	7.090	0.831	5.546	0.569	1.753	0.524
SDNet	6.682	0.661	5.059	0.501	1.562	0.450	5.428	0.474	3.054	0.642	1.111	0.410	7.013	0.729	5.514	0.500	1.544	0.525
TarDAL	7.163	0.800	4.789	0.484	1.670	0.412	6.478	0.699	3.140	0.628	1.526	0.420	7.126	0.812	4.140	0.510	1.450	0.407
UMFusion	6.699	0.673	3.710	0.516	1.677	0.409	5.761	0.488	2.442	0.597	1.077	0.299	6.881	0.771	3.420	0.546	1.618	0.470
DeFusion	6.724	0.712	2.996	0.493	1.592	0.325	5.950	0.759	2.855	0.589	1.339	0.471	6.634	0.740	3.027	0.513	1.366	0.412
ReCoNet	6.682	0.728	3.674	0.481	1.732	0.340	3.894	0.544	3.105	0.544	1.190	0.365	6.740	0.867	4.557	0.515	1.495	0.499
IGNet	7.099	0.764	5.247	0.521	1.756	0.459	6.124	0.762	3.290	0.655	1.562	0.485	7.140	0.882	5.615	0.575	1.762	0.539

Infrared image   Visible image   DIDFuse   U2Fusion   SDNet   TarDAL   UMFusion   DeFusion   ReCoNet   IGNet

Fig. 6. Detection visual comparisons of different fusion images on M³FD dataset. Our proposed IGNet can generate high-confidence detection results with visually appealing performance. The red, green, and yellow regions represent error, missing, and low-confidence detection, respectively. The blue areas denote our outstanding details.

DIDFuse, U2Fusion, SDNet, and ReCoNet appear some missing segmentation areas in the first sample. In addition to the segmentation results of IGNet, the "Color cone" in the second example cannot be accurately segmented. It is appropriate to mention that our proposed method can exploit cross-modality interaction features to efficiently segment the contours of labeled objects.

4.5.2 Quantitative Analysis. Table. 3 depicts the segmentation quantitative metric IoU for different categories, which presents IGNet outperforms other fusion methods in the segmentation task. Compared with the second-ranked method, our method improves mIoU in the ratio of 4.87%. For some infrared-sensitiveness labels, *e.g.*,

person, higher scores indicated that our method can more easily highlight thermal targets. Due to the high fidelity of fused images, the IoU of some visually appealing labels, *e.g.*, car and bike, still keeps high performance. Note that the proposed IGNet can also generate vivid fusion images while achieving accurate segmentation results.

4.6 Ablation Experiments

4.6.1 Study on Modules. The proposed SSM and GIM play a key role in improving the fusion effect. It is obvious that fusion results perform poorly in luminance without SSM as shown in Fig. 8. Also, the cross-modality features of infrared and visible branches can not

| Background | Car | Person | Bike | Curve | Car stop | Guardrail | Color cone | Bump |

Fig. 7. Segmentation visual comparisons of different fusion images on MFNet dataset. Our proposed IGNet can get the most accurate segmentation results compared to the ground truth. The red and green regions represent the error and missing segmentation, respectively.

Table 2

DETECTION QUANTITATIVE COMPARISONS OF OUR IGNET WITH SEVEN STATE-OF-THE-ART METHODS ON M^{3}FD DATASET. OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED IN RED AND BLUE, RESPECTIVELY.

| Method | AP@.5 | | | | | | mAP@.5 |
	People	Bus	Car	Motor	Truck	Lamp	
Infrared	0.807	0.782	0.888	0.640	0.652	0.703	0.745
Visible	0.708	0.780	0.911	0.702	0.697	0.865	0.777
DIDFuse	0.800	0.798	0.924	0.681	0.692	0.843	0.790
U2Fusion	0.793	0.785	0.916	0.663	0.710	0.872	0.789
SDNet	0.790	0.811	0.920	0.670	0.689	0.838	0.786
TarDAL	**0.817**	0.815	**0.948**	0.696	0.687	0.873	**0.806**
UMFusion	0.790	0.783	0.920	**0.728**	0.691	0.847	0.793
DeFusion	0.805	**0.827**	0.921	0.689	**0.714**	**0.876**	0.805
ReCoNet	0.792	0.784	0.915	0.693	0.698	**0.873**	0.792
IGNet	**0.816**	**0.824**	**0.928**	**0.730**	**0.721**	0.869	**0.815**

Table 3

SEGMENTATION QUANTITATIVE COMPARISONS OF OUR IGNET WITH SEVEN STATE-OF-THE-ART METHODS ON MFNET DATASET. OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED IN RED AND BLUE, RESPECTIVELY.

| Method | IoU | | | | | | | | | mIoU |
	Bac	Car	Per	Bik	Cur	C S	Gua	C C	Bum	
Infrared	0.821	0.663	0.592	0.513	0.347	0.398	0.422	0.414	0.479	0.516
Visible	0.899	0.774	0.482	0.586	0.372	0.517	0.451	0.432	0.506	0.558
DIDFuse	0.971	0.790	0.582	0.599	0.358	**0.526**	**0.619**	0.442	0.557	0.604
U2Fusion	0.974	0.817	**0.631**	**0.625**	0.408	0.523	0.520	0.448	**0.593**	**0.615**
SDNet	0.973	0.782	0.614	0.618	0.361	0.500	0.527	0.425	0.527	0.591
TarDAL	0.970	0.795	0.563	0.591	0.342	0.497	0.553	0.425	0.538	0.586
UMFusion	0.972	0.787	0.607	0.616	0.364	0.493	0.479	0.447	0.485	0.583
DeFusion	**0.975**	**0.820**	0.609	0.623	0.401	0.488	0.482	0.471	0.548	0.601
ReCoNet	0.973	0.813	0.598	0.610	**0.413**	0.519	0.544	**0.476**	0.552	0.610
IGNet	**0.976**	**0.838**	**0.639**	**0.667**	**0.435**	**0.532**	**0.626**	**0.511**	**0.586**	**0.645**

interact with each other without the decoration of the GIM, which causes the low contrast and halo artifacts of images. Furthermore, Fig. 9 reports the results of down-stream tasks. Due to the abundant semantic information extracted by the proposed module, the full modal can simultaneously obtain high-confidence detection and accurate segmentation results. The quantitative comparisons are depicted in Table. 4. It is not difficult to prove that the utilization of our proposed modules can bridge fusion and downstream tasks with a mutually beneficial situation.

4.6.2 Study on Leader Node. In order to avoid intermediate feature loss, we use leader nodes to guide the information delivery. Without the help of leader nodes, fused images often appear distorting in color. Meanwhile, some wrong regions may emerge in detection and segmentation results. In contrast, IGNet makes full use of feature maps delivered by the leader nodes inside graphs, enabling semantic information to be revealed in fused images. Fig. 10 performs the superiority of our proposed method on two different datasets.

4.6.3 Study on Parameters of Graph. We select one, three and five nodes to conduct each graph structure, aiming at verifying how the number of nodes N influence results. Except for the number of nodes, other parameters remain unchanged. It can be seen from

Fig. 8. Visual ablation comparisons of the SSM ($\mathcal{S}$) and GIM ($\mathcal{G}$) about fusion. The enlarged red and green circles are detailed patches of fusion results.

Table. 5 that when there is only a single node in a graph, the quantitative indicators perform undesirably. As the number increases to five, its performance is almost indistinguishable from our results (N = 3). However, the operating efficiency of the network will decrease with N rising. Considering this issue, we employ three nodes in each graph, which can balance the quality of images and inference speed. Similarly, the number of loop L are preset to three for a trade-off.

Table 4

QUANTITATIVE ABLATION RESULTS OF MODULES ON TWO DIFFERENT DATASETS. OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED IN RED AND BLUE, RESPECTIVELY.

Model	$\mathcal{S}$	$\mathcal{G}$	Dataset:M³FD							Dataset:MFNet						
			EN	VIF	AG	CC	SCD	$Q_{ab/f}$	mAP@.5	EN	VIF	AG	CC	SCD	$Q_{ab/f}$	MIoU
M1	✗	✔	7.052	0.873	5.608	0.570	1.699	0.504	0.801	6.045	0.751	3.187	0.633	1.549	0.470	0.624
M2	✔	✗	7.009	0.735	5.536	0.507	1.574	0.515	0.791	5.443	0.498	3.091	0.642	1.551	0.426	0.587
M3	✔	✔	7.140	0.882	5.615	0.575	1.762	0.539	0.815	6.124	0.762	3.290	0.655	1.562	0.485	0.645

Fig. 9. Visual ablation comparisons of the SSM ($\mathcal{S}$) and GIM ($\mathcal{G}$) about down-stream tasks. The detailed regions are marked.

Table 5

QUANTITATIVE ABLATION RESULTS ABOUT THE NUMBER OF NODES (N) AND LOOPS(L) IN THE GRAPH. OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED IN RED AND BLUE, RESPECTIVELY.

Model	N	L	EN	VIF	AG	CC	SCD	$Q_{ab/f}$
M1	1	3	6.032	0.751	3.289	0.647	1.553	0.480
M2	3	3	6.124	0.762	3.290	0.655	1.562	0.485
M3	5	3	6.125	0.762	3.289	0.657	1.563	0.485
M1	3	1	6.111	0.744	3.277	0.641	1.559	0.476
M2	3	3	6.124	0.762	3.290	0.655	1.562	0.485
M3	3	5	6.124	0.764	0.291	0.655	1.564	0.487

5 CONCLUSION

In this paper, an interactive cross-modality framework based on graph neural network was proposed for infrared and visible image fusion. We presented a graph interaction module to learn mutual features from different branches, which can emphasize outstanding textures in source images. Aiming at preventing information from missing, the leader nodes were proposed to guide the feature propagation between adjacent graphs. In addition, abundant semantic information was also extracted by our proposed method, thus we could achieve well-performance detection and segmentation results. Extensive experiments proved our method is advanced in IVIF and down-stream tasks.

In the future, we tend to bridge multi-modality fusion, target detection, and image segmentation in a unified framework. In other words, it is worth further exploiting how to generate a fusion image that can also perform well in detection and segmentation tasks.

Source images   Infrared   Visible   w/o g_i^* Ours

Fig. 10. Visual ablation comparisons of the leader nodes (g_i^*) about fusion, detection and segmentation. With the help of leader nodes, the image details and down-stream results can perform more vividly and accurately. The enlarged red and green boxes are detailed patches of corresponding results.

ACKNOWLEDGMENTS

REFERENCES

[1] V Aslantas and Emre Bendes. 2015. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications* 69, 12 (2015), 1890–1896.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[3] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. 2017. 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, 5108–5115.

[4] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. 2013. A new image fusion performance metric based on visual information fidelity. *Information fusion* 14, 2 (2013), 127–135.

[5] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[6] Huimin Huang, Lanfen Lin, Yue Zhang, Yingying Xu, Jing Zheng, XiongWei Mao, Xiaohan Qian, Zhiyi Peng, Jianying Zhou, Yen-Wei Chen, et al. 2021. Graph-bas3net: Boundary-aware semi-supervised segmentation network with bilateral graph convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7386–7395.

[7] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. ReCoNet: Recurrent Correction Network for Fast and Efficient Multi-modality Image Fusion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Springer, 539–555.

[8] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. 2022. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3783–3791.

[9] Jia Lei, Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K Kasabov. 2023. GALFusion: Multi-exposure Image Fusion via a Global-local Aggregation Learning Network. *IEEE Transactions on Instrumentation and Measurement* (2023).

[10] Hui Li and Xiao-Jun Wu. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing* 28, 5 (2018), 2614–2623.

[11] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. 2020. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia* 23 (2020), 1383–1396.

[12] Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K Kasabov. 2022. Learning a coordinated network for detail-refinement multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[13] Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K Kasabov. 2023. GeSeNet: A General Semantic-Guided Network With Couple Mask Ensemble for Medical Image Fusion. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[14] Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K Kasabov. 2023. Infrared and visible image fusion based on residual dense network and gradient loss. *Infrared Physics & Technology* 128 (2023), 104486.

[15] Tengpeng Li, Kaihua Zhang, Shiwen Shen, Bo Liu, Qingshan Liu, and Zhu Li. 2021. Image co-saliency detection and instance co-segmentation using attention graph clustering based graph convolutional network. *IEEE Transactions on Multimedia* 24 (2021), 492–505.

[16] Yao Li, Xueyang Fu, and Zheng-Jun Zha. 2021. Cross-patch graph convolutional network for image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4651–4660.

[17] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. 2022. Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Springer, 719–735.

[18] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5802–5811.

[19] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. 2021. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 1 (2021), 105–119.

[20] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. 2023. HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion* 95 (2023), 237–249.

[21] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. 2021. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters* 28 (2021), 1818–1822.

[22] Jinyuan Liu, Yuhui Wu, Guanyao Wu, Risheng Liu, and Xin Fan. 2022. Learn to Search a Lightweight Architecture for Target-Aware Infrared and Visible Image Fusion. *IEEE Signal Processing Letters* 29 (2022), 1614–1618.

[23] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. 2020. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing* 30 (2020), 1261–1274.

[24] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. 2020. Cascade graph neural networks for RGB-D salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 346–364.

[25] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica* 9, 7 (2022), 1200–1217.

[26] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. 2020. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing* 29 (2020), 4980–4995.

[27] Nirmala Paramanandham and Kishore Rajendiran. 2018. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. *Infrared Physics & Technology* 88 (2018), 13–22.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[29] Parul Shah, Shabbir N Merchant, and Uday B Desai. 2013. Multifocus and multi-spectral image fusion based on pixel significance using multiresolution decomposition. *Signal, Image and Video Processing* 7 (2013), 95–109.

[30] Kechen Song, Liming Huang, Aojun Gong, and Yunhui Yan. 2022. Multiple graph affinity interactive network and a variable illumination dataset for RGBT image salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[31] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4003–4011.

[32] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 10 (2022), 6700–6713.

[33] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. 2023. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion* 91 (2023), 477–493.

[34] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion* 82 (2022), 28–42.

[35] Alexander Toet. 2017. The TNO multiband image data collection. *Data in brief* 15 (2017), 249–251.

[36] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. 2022. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876* (2022).

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[38] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. 2021. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5475–5484.

[39] Han Xu, Meiqi Gong, Xin Tian, Jun Huang, and Jiayi Ma. 2022. CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Computer Vision and Image Understanding* 218 (2022), 103407.

[40] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1 (2020), 502–518.

[41] Costas S Xydeas, Vladimir Petrovic, et al. 2000. Objective image fusion performance measure. *Electronics letters* 36, 4 (2000), 308–309.

[42] Hao Zhang and Jiayi Ma. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision* 129 (2021), 2761–2785.

[43] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion* 76 (2021), 323–336.

[44] Jun Zhang, Licheng Jiao, Wenping Ma, Fang Liu, Xu Liu, Lingling Li, Puhua Chen, and Shuyuan Yang. 2023. Transformer based Conditional GAN for Multimodal Image Fusion. *IEEE Transactions on Multimedia* (2023).

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881–2890.

[46] Zixiang Zhao and Haowen et al. Bai. 2022. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. *arXiv preprint arXiv:2211.14461* (2022).

[47] Zixiang Zhao and Shuang et al. Xu. 2020. DIDFuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210* (2020).