IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 5 July 2024, accepted 11 July 2024, date of publication 15 July 2024, date of current version 26 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428407

RESEARCH ARTICLE

Infrared and Visible Image Fusion via General Feature Embedding From CLIP and DINOv2

YICHUANG LUO[1], **FANG WANG**[1], **AND XIAOHU LIU**[2]
[1]Department of Intelligent Science and Engineering, Xi'an Peihua University, Xi'an 710125, China
[2]Trine Engineering Institute, Shaanxi University of Technology, Hanzhong 723001, China

Corresponding author: Xiaohu Liu (lxh@snut.edu.cn)

This work was supported in part by Xi'an Peihua University under Grant PHKT2307.

ABSTRACT Jointing multi-modal image fusion and subsequent high-level tasks is attracting more researches to achieve both mutual promotions. However, owing the feature gap between the two tasks, complicated network structure and training strategies need to be redesigned for specific different datasets. To address these issues, this paper proposes an infrared and visible image fusion via general feature embedding from frozen CLIP and DINOv2 models. The core idea is that the general semantic features from CLIP model are injected into the fusion network with the DINOv2-based segmenter as a constraint. Specially, the feature merging module and injection strategies are design to generate the semantic features that are compatible with fusion features meanwhile aligned with DINOv2 features. Leveraging the generalization ability of these foundation models, the proposed network can be optimized mutually to promote the training process. Comprehensive experiments on the four public datasets demonstrate the effectiveness of our method.

INDEX TERMS CLIP, DINOv2, feature alignment, image fusion, multi-modal fusion, semantic segmentation.

I. INTRODUCTION

The infrared and visible image fusion technology largely promotes the real applications, in which the visible images are used to catch texture details and the infrared images to supply robust object outlines without being affected by light. Mainly, there are two categories: perception-oriented methods and semantic-driven methods [1] or joint learning methods [7]. The perception-oriented methods focus on pixel-level fusion for better visual effects, such as sparse representations [2], saliency analysis [3], adversarial training strategy [4] and etc. To facilitate the subsequent high-level tasks, the semantic-driven methods reinforce semantic information in fused images by taking the high-level task model as a supervision, either utilizing high-level models as a constrain by task-specific loss, e.g., the SeAFusion [5] cascades the segmentation model behind the fusion network, or designing the feature-level fusion modules to inject the semantic features from the high-level tasks, e.g., DetFusion [6] utilizes object-level features learned from detection model to guide the fusion.

Although the semantic-driven methods achieve satisfactory fusion results, these methods typically deploy well-established feature extraction network to extract semantic features from source images. Subsequently, specific fusion modules are devised to integrate complementary features based on the feature extraction network, and task-specific prediction networks are applied to accomplish the desired tasks. Therefore, most methods focus on designing networks [6], [8], [9] and introducing constraints [5], [10], [11], as shown in Fig.2 (a), (b). And many efforts on specific multi-stage training strategy need to be taken, e.g., the MetaFusion [7] proposed a mutual promotion learning between fusion and detection task, and training process contained four steps: fusion pre-training, detection fine-tuning, feature transformation and meta-feature generation, and mutual promotion.

Additionally, the existing joint learning methods are tailored solely for specific task datasets, e.g., M^3FD [23], which fail to be generalized to other datasets, e.g., MFNet [24]. These methods utilize specific high-level models to constrain

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

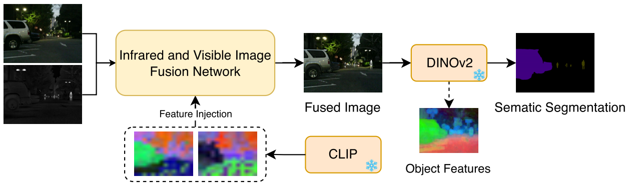

FIGURE 1. An overview of the proposed GFFusion that fuses the semantic features to promote semantic segmentation based on vision foundation models.

the fusion results, which may limit the generalization of the fused images to other tasks. And the feature extraction network is usually unable to adapt effectively to the domain variation between infrared and visible images, which leads to performance degradation [1]. Furthermore, with the vision model evolving rapidly, models with strong generalization capability, such as CLIP [12] and DINO [13], are verified on unimodal datasets. But the potential on multi-modality fusion has not been excavated. Therefore, we leverage the general features from CLIP to guide the infrared and visible image fusion. Unfortunately, CLIP features are not well performed for subsequent tasks [14]. To address this issue, we propose a feature injection module, and is constrained by the DINOv2-based high-level task, as shown in Fig.1.

Specifically, the infrared and visible image fusion via General Feature Embedding from CLIP and DINOv2 is proposed, named GFFusion, and an overview of the proposed method is shown in Fig.1 and Fig.2 (c). GFFusion consists of infrared and visible image fusion network (FuNet), semantic segmentation network (SSNet) based on DINOv2, and multi-level semantic feature injection network (FINet) from CLIP. In particular, benefitting of the scalability and generalization ability of the vision foundation models, the FuNet, FINet optimization and SSNet fine-tuning can be implemented at the same time with the segmentation loss and fusion loss, where the alternate optimization steps are not needed. The motivation is the analysis results from [14], where the features of CLIP can exhibit biases towards local patterns, which contain low-level detailed information, while the DINOv2 can capture fine-grained localization information, which is beneficial for positioning ability. Specially, the semantic features of CLIP and DINOv2 are compatible, which can be aligned with a MLP layer. Further, in the optimization process, both the CLIP and DINOv2 models are stay frozen. And the SSNet can be fine-tuned only using a light-weight segmentation head, such as linear layer. For FINet, different injection strategies are proposed that integrates the different layer features of CLIP to inject semantic information into the fusion network.

The main contributions can be summarized as follows: (i) we explore the different joint learning framework of infrared and visible image fusion and high-level tasks, as presented in Fig.2. And GFFusion is proposed to obtained superior performance on fusion and semantic segmentation. (ii) We inject the semantic features from CLIP into the fusion network with different strategies, to implicitly align the semantic features from fusion result with the high-level DINOv2 features. (iii) Sequentially, the jointly training strategy is introduced to mutually promote the proposed FuNet, FINet and SSNet learning, as shown in Alg.1. And extensive experiments demonstrate the superiority of our proposed method on image fusion and semantic segmentation.

The remainder of this paper is organized as follows. In Section II, we briefly introduce the related works of image fusion, semantic-driven fusion, and vision foundation models. In Section III, we elaborate on the proposed GFFusion, including the overall framework and each module design. Section IV illustrates the performance of our method in comparison with others, and the ablation study. Section V concludes this paper.

II. RELATED WORKS

A. INFRARED AND VISIBLE IMAGE FUSION

The infrared and visible image can provide complementary information for each other to promote the subsequential tasks. Before the deep learning era, the fusion methods, such as sparse representation [2], and low-rank representation [15], are proposed, but cannot tackle complex scenes well. Nowadays, deep learning-based methods [5], [6], [16] are raised, and specially, the feature-based fusion methods became the main-stream. Tang et al. [8] propose a fusion method with cross-domain long-range learning based on Swin Transformer architecture. Xu et al. [17] use feature extraction and measurement to estimate the degree of information preservation in image fusion. However, most of them ignore the gap between fusion result and the high-level tasks, resulting to the performance degradation on subsequential tasks.

B. SEMANTIC-DRIVEN FUSION

To facilitate the subsequent high-level tasks, the semantic-driven methods are proposed to reinforce the semantic information in fused images. These methods either utilize high-level models as a constrain by task-specific loss, e.g., the SeAFusion [5] cascades the segmentation model behind the fusion network, or design the feature-level fusion modules to inject the semantic features from the high-level tasks, e.g., DetFusion [6] utilizes object-level features learned from detection model to guide the fusion. Although the semantic-driven methods achieve satisfactory fusion results, these methods typically deploy well-established network and complex alternate optimization procedure, such as [6] and [7]. In particular, for different datasets, the whole process needs to be repeated. On the contrary, we leverage the scalability and generalization ability of different vision foundation models to guide the fusion network without a complicated optimization design.

C. VISION FOUNDATION MODEL

More recently, some models, that are trained at large scale data in an unsupervised manner and capable of being

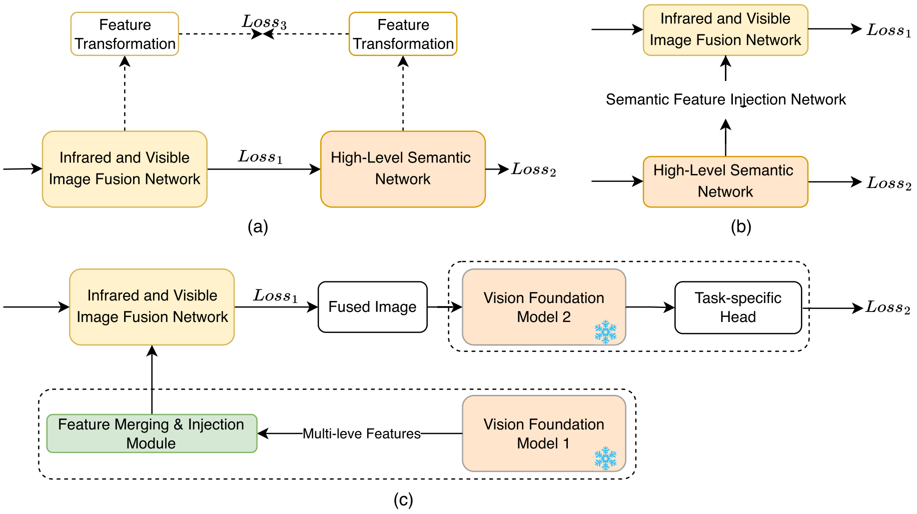
Feature Transformation
$Loss_3$
Feature Transformation
Infrared and Visible Image Fusion Network
$Loss_1$
High-Level Semantic Network
$Loss_2$
(a)
Infrared and Visible Image Fusion Network
$Loss_1$
Semantic Feature Injection Network
High-Level Semantic Network
$Loss_2$
(b)
Infrared and Visible Image Fusion Network
$Loss_1$
Fused Image
Vision Foundation Model 2
Task-specific Head
$Loss_2$
Feature Merging & Injection Module
Multi-leve Features
Vision Foundation Model 1
(c)

FIGURE 2. Different semantic injection methods of infrared and visible image fusion (FuNet) and high-level semantic tasks, such as segmentation (SSNet). (a) Cascading learning method. The FuNet treats SSNet as a constraint by loss function as showed with solid line, the other proposes the transformed feature embedding to bridge the semantic gap as showed with dash line. (b) Parallel learning method. The FuNet fuses the semantic information from the optimized SSNet by the design of semantic feature inject network (FINet). (c) Hybrid learning method. The FuNet fuses the semantic information from pre-trained frozen vision foundation model with a proposed FINet, which is optimized by the cascading high-level task with another frozen vision foundation model.

generalized (e.g., fine-tuned) to a wide range of down-stream tasks, are emerged, denoted as foundation model. For vision, such large-scale pre-training methods such as CLIP [12], which learn directly from large-scale image-text pairs, show very encouraging progress for efficient transfer learning and zero-shot capability. And these methods focus on contrastive learning [12], [13], [36] and masked image modeling [18]. Specially, for contrastive learning, DINOv2 [13] pretrains the image encoder on large image data, which shows a superior understanding of object parts and scene geometry across image domains. Image-text contrastive learning as CLIP [12] employs the natural language as weak supervision to guide the learning of visual features. For masked image modeling, MAE [18] proposes a masked autoencoder for reconstructing image pixels. Inspired by above methods, we design the semantic feature injection module from CLIP, which constrained by the DINOv2 based segmentation task, to fully leverage the capabilities of vision foundation model, addressing the problems presented in the introduction section and reducing the design complexity.

III. PROPOSED METHOD

In this section, we first summarize existing semantic-driven paradigms, then the proposed method is introduced, that integrates CLIP and DINOv2 with multi-level features injection to enhance the generalization and semantic capabilities.

A. HYBRID SEMANTIC-DRIVEN LEARNING METHOD

Existing semantic-driven learning methods can be divided into two categories: the cascading learning method and parallel learning method, as shown in Fig.2 (a), (b), where $Loss_1$ represents the fusion loss, $Loss_2$ denotes the task-related loss, and $Loss_3$ is the similarity metric loss. Detailed as follows:

1) CASCADING LEARNING METHOD

This method cascades the FuNet ψ with SSNet ϕ, leveraging the semantic loss to feed high-level semantic information back to the image fusion network, such as SeAFusion [5], as shown in Fig.2 (a) with solid lines. However, directly utilizing the SSNet constraint to guide the FuNet results in a limited effect [7], owing to the mismatching between SSNet features and FuNet features. To address this problem, MetaFusion [7] proposed a meta-feature embedding network for feature alignment, as shown in Fig.2 (a) with dash lines. This process can be described as (1):

$$\begin{aligned} &min_{\theta_f} L_f(\psi) + L_s(\phi) \\ &s.t.\ I_{fus} = \psi(I_{rgb}, I_{ir}; \theta_f) \\ &\quad I_{seg} = \phi(I_{fus}; \theta_s). \end{aligned} \tag{1}$$

In which, I_{rgb} and I_{ir} represents the visible image and infrared image respectively. L_f is fusion loss, L_s is segmentation loss. I_{fus} and I_{seg} denote the fusion result and segmentation result respectively. θ_s and θ_f are the parameters of SSNet and FuNet. On optimization, an iterative strategy is need to first train the SSNet, then fixing SSNet to optimize the FuNet for several eposes.

2) PARALLEL LEARNING METHOD

In cascading learning method, there is no explicit semantic information from SSNet injected into FuNet, which are learned by SSNet to guide the optimization of FuNet.

While for the parallel learning method, an explicit semantic information injection module FINet χ is designed with the FuNet, where the FuNet and FINet need to be optimized together. Specially, the SSNet, extracting sufficient semantic features, provides a prior information for FuNet to fulfill the semantic requirements for high-level vision tasks, such as DetFusion [6], as shown in Fig.2 (b), which can be formulated as (2) and (3):

$$\begin{aligned} & min_{\theta_s} L_s(\phi) \\ & s.t. F_s = \phi(I_{rgb}, I_{ir}; \theta_s). \end{aligned} \tag{2}$$

$$\begin{aligned} & min_{\theta_f} L_f(\psi) \\ & s.t. I_{fus} = \psi(I_{rgb}, I_{ir}; \chi(F_s; \theta_i; \theta_f). \end{aligned} \tag{3}$$

where F_s represents the multi-level features from SSNet, and θ_i is the parameters of FINet. For network training, the same iterative optimization strategy is needed. Specially, the FINet usually leverage the cross-attention or concatenation mechanisms to inject the semantic information to FuNet, for example, the two branches of visible and infrared cross-attention are applied in DetFusion.

3) HYBRID LEARNING METHOD

As mentioned in the introduction section, the above methods are specific for certain datasets and need complicated training strategies. Take advantage of the generalization abilities of the vision foundation models, we proposed the hybrid learning method, which integrates the cascading learning method and parallel learning method and consists of explicit semantic information and high-level task constraint. Specifically, the hybrid learning method constrains the FuNet to guide the fusion process with SSNet, meanwhile injects semantic information to FuNet using FINet, as shown in Fig.2(c). This method can be formulated as (4):

$$\begin{aligned} & min_{\theta_f} L_f(\psi) + L_s(\phi) \\ & s.t. I_{fus} = \psi(I_{rgb}, I_{ir}, \chi(F_m; \theta_i); \theta_f) \\ & I_{seg} = \phi(I_{fus}; \theta_s). \end{aligned} \tag{4}$$

In which, F_m denotes the multi-level features from frozen foundation model. Specially, motivated by [14], DINOv2 shows superior understanding of object parts across image domain and can capture fine-grained localization information. CLIP vision model contains more information regarding local objects, such as shape or texture, as shown in Fig.3. In which, the first row is the visible image and second row is the infrared image. Meanwhile, it can be observed that the pre-trained models show strong generalization ability for both modalities. Therefore, in our method, the CLIP vision model is utilized for the input of FINet, and the DINOv2 model is taken as the backbone of SSNet.

B. ARCHITECTURE

The overall architecture is illustrated in Fig.4. GFFusion integrates CLIP (based on ViT-Base) with multi-level feature merging to enhance the fusion performance, and optimizes the parameters with DINOv2-based segmenter (based on ViT-Base) as a constraint to simplify the training process and speed up convergence. Specifically, we adopt the convolutional neural network for FuNet to balance the performance and efficiency, and the gradient residual dense block (GRDB) from [5] is applied to enhance the fine-grained spatial details.

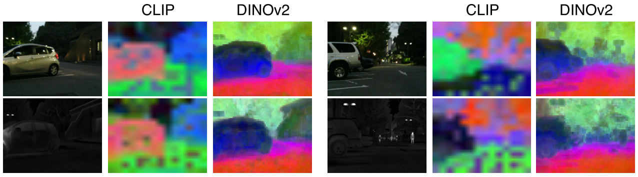

FIGURE 3. The general features extracted from the CLIP and DINOv2 models. The backbone of the two models is the same ViT-B/16 [19], and the PCA algorithm is used for visualization.

1) FEATURE EXTRACTION OF FUNET

For inputs I_{rgb} and I_{ir}, the Conv operations are used to extract the fine-grained spatial features:

$$F_{cls} = GRDB(Conv(I_{cls})), cls \in \{rgb, ir\}. \tag{5}$$

where the $F_{cls} \in \mathbb{R}^{H' \times W' \times 48}$ denotes the features of input images, the $Conv$ represents the 3×3 convolution with Leaky ReLU [20] as activation function to output the features with embedding dimension 16, the and the H' and W' is the feature resolution. The GRDB is consist of two branches: one dense connection branch with two concatenated convolutions, the other gradient branch with gradient operator. The two branch features are integrated with element-wise addition, where the 1×1 convolutions are used to align the channel dimension. After the concatenation of visible and infrared features, we can obtain the preliminary fusion features z_f with dimension 96.

2) MULTI-LEVEL SEMANTIC FEATURES

Specifically, following [14], denote the visual encoder of CLIP as π. Given the inputs I_{rgb} and I_{ir}, the patch token features are extracted by all layers of CLIP as $\pi(I_{rgb}) = [f^1_{rgb}, \cdots, f^l_{rgb}, \cdots, f^{11}_{rgb}]$, where $f^l_{rgb} \in \mathbb{R}^{196 \times 768}$. Similarly, the features of I_{ir} are $\pi(I_{ir}) = [f^1_{ir}, \cdots, f^l_{ir}, \cdots, f^{11}_{ir}]$. Then we integrate these features by element-wise maximization operation by (6):

$$f_l = max_{\odot}(f^l_{rgb}, f^l_{ir}), l \in \{1, 2, \cdots, L\}. \tag{6}$$

where L is the total number of layers, here, $L = 11$. The intuition is that the features from different modality can provide complementary information for each other by maximization operation, meanwhile maintain the efficiency.

3) FEATURE MERGING

To effectively integrate the shallow and deep features, several feature merging strategies for combing the multi-level features are explored, as shown in Fig.4. Detailed as:

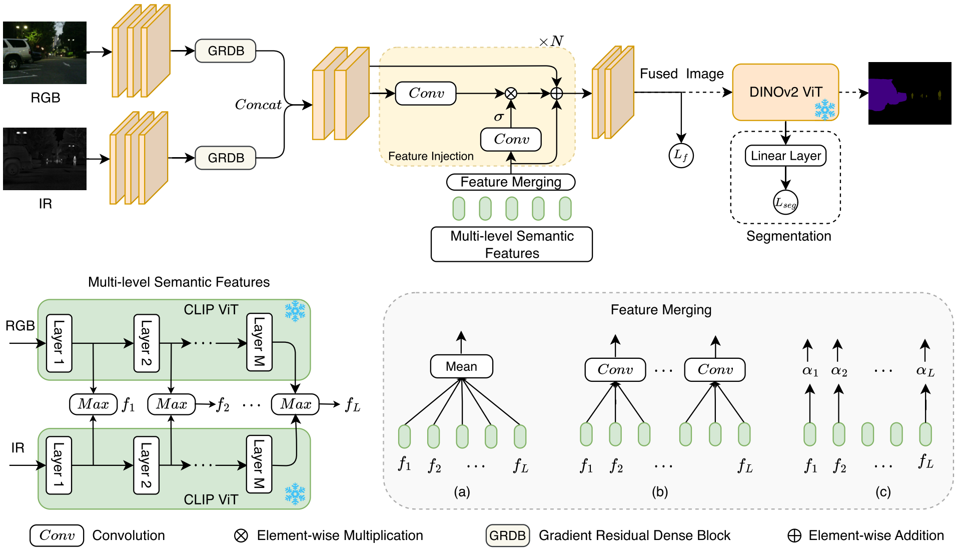

FIGURE 4. An illustration of our proposed architecture. The visible and infrared images are input to CLIP, and the features from two modalities are incorporated by max operator. Further the multi-level features are fused by feature merging, and the fused features are injected into FuNet with feature injection module. The DINOv2 is used in SSNet for semantic segmentation.

- Mean. averaging all the output features as $z_c = (f_1 + f_2 + \cdots + f_L)/L$.
- Group-Conv. using a group convolution [21] and batch normalization to integrate the features and then summed by learnable weights as $z_c = \omega_1 GroupConv(f_1, \cdots, f_{L/N}) + \cdots + \omega_N GroupConv(f_{(N-1)L/N}, \cdots, f_L)$. In which g denotes the g-th group, N is the number of groups, in our setting, N=2. And ω refers to the weights and are summed up to 1.
- Weighted-Sum. weighted summing the features with a learnable parameter as $z_c = \omega_1 f_1 + \omega_2 f_2 + \cdots + \omega_L f_L$.

4) FEATURE INJECTION

A gated linear unit (GLU) [21] is leverage to implement a gating mechanism over the features of z_f and z_c as (7):

$$z_g = z_f \otimes \sigma(upsampling(Conv(z_c))). \tag{7}$$

where 1×1 convolution is applied on z_c to stay same channel dimension with z_f, then upsampling operation is used to align the spatial dimension. $\otimes$ is the point-wise multiplication, and σ denotes the sigmoid function. In this way, the semantic information is only utilized as a guidance, to further facilitate the subsequent task, we add the linear mappings to project between z_c and z_g, then implement the element-wise addition with (8):

$$z = z_g \oplus Linear(z_c). \tag{8}$$

where the linear projection can also leverage the alignment between the CLIP features and DINOv2 features, as analyzed in [14]. To further stabilize the training process, the residual connections from the z_f to the FINet are added. And the fusion image I_{fus} can be obtain with another convolution layers performing batch normalization on feature embedding z.

5) SEMANTIC SEGMENTATION

Benefit from the zero-shot abilities of the DINOv2, a simple linear layer is trained to predict class logits from the patch tokens, as shown in Fig.4. Then the logits map is upsampled to the output resolution to obtain the final segmentation map. Specifically, following [13], let I_{fus} be the input fused image, the patch tokens of DINOv2 $f_d \in \mathbb{R}^{196 \times 768}$. And the segmentation map can be obtained by (9):

$$I_{seg} = upsampling(\sigma(Linear(f_d))). \tag{9}$$

Alternatively, the pre-trained DINOv2-based segmenter [13] on ADE20K can be used as initialization.

6) LOSS FUNCTION

To boost the fusion quality and subsequent task performance at the same time by injecting semantic information into fusion image, our loss function consists of two aspects: one is structure and texture loss L_{st} to maintain the visual fidelity, the other is the semantic segmentation loss L_{ss} to make sure the contribution of the fusion image to high-level task. Specifically, the L_{st} contains the structural similarity index (SSIM) [22] and the texture loss [5], defines as (10):

$$\begin{aligned} L_{st} = {} & (\frac{1 - SSIM_{I_{fus},I_{rgb}}}{2} + \frac{1 - SSIM_{I_{fus},I_{ir}}}{2}) \\ & + \beta/(HW) \left| \left| \nabla I_{fus} \right| - max(\left| \nabla I_{improve} \right|, \left| \nabla I_{rgb} \right|) \right| . \end{aligned} \tag{10}$$

where H and W denotes the image resolution, β denotes the balancing coefficient between these two losses. And ∇ represents the Sobel gradient operator, $|\cdot|$ is $L1$ norm. The segmentation loss is the cross-entropy loss between the predicted segmentation results I_{seg} and ground truth labels I_{gt}.

Particularly,

$$L_{ss} = -\sum_{c} I_{gt} log(I_{seg}), c \in \{1, 2, \cdots, C\}. \tag{11}$$

In which, C denotes the number of classes. The final loss can be defined as (12):

$$L_{total} = L_{st} + \lambda L_{ss}. \tag{12}$$

where, λ is used to adjust the semantic segmentation loss.

```
Algorithm 1 GFFusion Training
Input: Visible images $I_{rgb}$ and infrared images $I_{ir}$
Output: Fusion images $I_{fus}$
  Load pre-trained CLIP visual encoder weights and DINOv2-based segmenter.
  while not converged do
    Sample image pairs $(I^p_{rgb}, I^p_{ir})$ from Input
    Update the parameters $\theta_f$ and $\theta_i$ of the network by Adam optimizer according to Eq.(12): $\nabla_\theta(L_{total})$
    if epochs > $q$ then
      Increase $\lambda$ with a cosine scheduler.
      Decrease the learning rate by specific decaying ratio.

      Update the sematic parameter $\theta_s$ by Adam optimizer according Eq.(12): $\nabla_{\theta_s}(L_{total})$
    end if
  end while
```

C. TRAINING

Leveraging the generalization ability of the pre-trained vision foundation model, we can train the whole networks jointly according to Eq.(12), which is analyzed in the ablation study, under the situation of ground truth missing of fusion images. Particularly, after a few iterations, we increase the weight λ to make the SSNet guide the FuNet and FINet more precisely. The training process of our method is shown in Alg.1, where the CLIP model and DINOv2 model stay frozen. And the Adam optimizer is applied to update the parameters.

IV. EXPERIMENTS

A. SETUP

1) DATASET

We conduct the experiments on four widely-used datasets: M^3FD [23], MFNet [24], RoadScene [17] and TNO [26]. Where, the image pairs of RoadScene and TNO datasets are only used for testing. Besides, MFNet is adopted to evaluate semantic segmentation performance, and the image pairs for object detection task in M3FD and MFNet are transformed to segmentation masks based on the DINOv2 features with PCA, as shown in Fig.5.

2) IMPLEMENTATION

Our framework is implemented with PyTorch on a NVIDIA GeForce RTX 4090 GPU 24G. The FuNet and FINet are trained using Adam with learning rate 1×10^{-3}, respectively. And the segmentation head of SSNet is trained using learning rate 1×10^{-4} with 0.1 decaying rate every 10 epochs. We firstly train the network for 100 epochs. Then, we fine-tune the SSNet for 50 epochs, meaning the q is set to 50. The hyperparameter λ is set to 0.2, and N is set to 1.

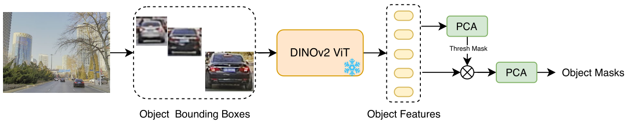

FIGURE 5. An overview of the object masks obtaining based on the bounding boxes. Firstly, the PCA is used to get the first component, then the foreground mask is obtained by thresholding. Based on the foreground feature, again the PCA is adopt to get foreground object mask.

TABLE 1. Comparison of different semantic-driven fusion methods on *M^3FD*.

Method	Model	Metric		
		SCD	EN	VIF
Cascading learning method	SeAFusion [5]	1.586	6.846	0.722
Parallel learning method	PSFusion [1]	1.832	7.400	0.824
Hybrid learning method	GFFusion (Ours)	1.872	7.641	0.852

3) METRIC

Following [7], Three metrics are used for fusion quality evaluation: entropy (EN) [27], sum of the correlations of differences (SCD) [28] and visual information fidelity (VIF) [29]. EN evaluates the information richness in an image, and the higher EN means more information. SCD evaluates the correlation between the input images and fused image. The higher MI illustrates more information of the input images is fused. VIF measures the ability to extract visible information from the input image, and a larger VIF represents less visible distortion in the fused result. Moreover, we use mIOU to comprehensively evaluate semantic segmentation performance. A higher mIOU means better segmentation effect.

B. ABLATION STUDIES

Effect of hybrid learning. In Section III-A, we summarized several different learning paradigms for semantic-driven image fusion, as shown in Fig.2. Then the hybrid learning method is introduced that help the fusion network fuse the semantic information from vision foundation models. The comparison results among these methods are shown in Table.1. Our GFFusion achieves comparable results on image fusion. The reason is that vision found model itself has strong generalization for high-level tasks, such as classification, segmentation and depth estimation, and the injected semantic features can be easily aligned to eliminate the mismatch of the other learning methods.

Study of the semantic feature merging strategy. We implement multi-level feature merging to inject the general semantic feature into FuNet, as describe in Section III-B.

TABLE 2. Study of the semantic feature merging by comparison different strategies on M^3FD.

Method	Metric		
	SCD	EN	VIF
Mean	1.653	7.140	0.762
Group-Conv	1.729	7.351	0.821
Weighted-Sum	1.872	7.641	0.852

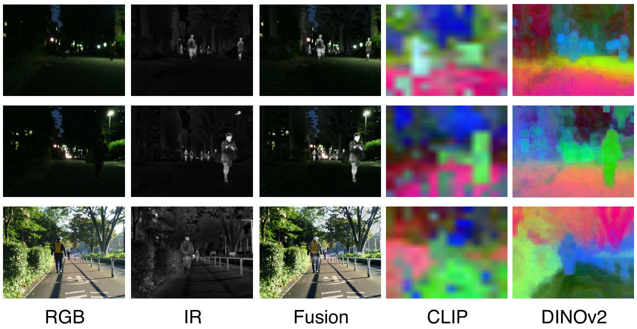

FIGURE 6. Feature visualization by applying PCA on the features of fusion images extracted by CLIP and DINOv2. Where the visualization of CLIP is upsampled with bilinear interpolation to maintain the same resolution with DINOv2, and the images are sampled from MFNet.

Here, we study the different strategies to show the results in Table.2. It can be observed that the Mean strategy simply averaging all the features of CLIP is hard to achieve a satisfactory result, and the Weighted-Sum strategy further improves the performance. With the subsequent linear embedding feature injection, a higher image fusion quality can be obtained. Since the combination of Weighted-Sum and linear embedding injection provide more semantic features to fusion network, which implicitly aligned with the DINO features, which is consistent with analysis of [14].

Effect of CLIP and DINOv2 features. Owing that the CLIP model is trained with weakly-supervised image-text pairs, it primarily learns image-level features, and inadequately explore the grounding details of local object parts, as depicted in Fig.3. And the visual features obtained from the DINOv2 contain more detailed information regarding local objects, and can capture the fine-grained localization information. To verify this effect, we further visualize the features extracted from the fusion images, as shown in Fig.6. So, the fusion network can achieve more fusion performance by injection the semantic information from CLIP, meanwhile achieve comparable segmentation performance benefiting the fine-grained localization information captured by DINOv2.

Effect of the DINOv2-based segmentation. Here, we study the generalization ability of the DINOv2-based semantic segmentation to support the training strategy for GFFusion, as shown in Fig.7. It can be seen that the DINOv2-based segmenter can achieve satisfactory results without fine-tuning, thus, it can be used as a constraint for fusion network at the beginning. And with the optimization process, it can reach a better segmentation performance on fusion images.

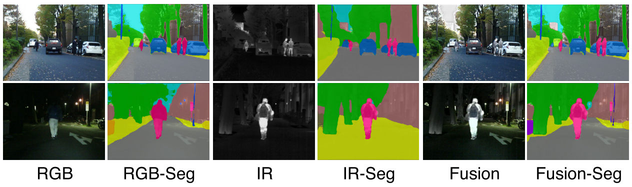

FIGURE 7. The segmentation results on visible images, infrared images and fusion images based on DINOv2 without fine-tuning on MFNet.

C. COMPARISON WITH SOTAS

1) THE FUSION RESULTS

In Table.3, compared with other state-of-the art fusion methods, our proposed method, GFFusion, achieves comparable results, which shows that the GFFusion can preserve the features from pixel-level (EN and SCD) and semantic-level at the same time. Specifically, it can be observed that compared with other semantic-driven methods, such as SeAFusion and PSFusion, the semantic-driven methods is superior to other methods.

Further, the qualitative results of the proposed GFFusion with several fusion methods: IFCNN [30], UMF-CMGR [31], SwinFusion [8] and SeAFusion [5], are shown in Fig.8. It can be observed that IFCNN and UMF-CMGR produce low contrast objects, and SwinFusion and SeAFusion generate smooth-effect edges. While the fusion images produced by GFFusion contain more edge details and high contrast objects.

More comparisons are implemented on M^3FD, RoadScene and TNO datasets, as shown in Fig.9. Our method can fuse both features of the visible and infrared images. Specially, for the strong light situation, as shown in the second row, the proposed method can fuse the person and car object with clear edges and details.

2) THE SEGMENTATION RESULTS

We provide quantitative results of different segmentation methods in Table.4, where the segmentation head of DINOv2-Seg is optimized on the dataset. Our GFFusion generally achieves the comparable performances. In detail, the semantic-driven methods, such as SeAFuson and PSFusion, which fused more semantic features, can achieve high segmentation performance with our SSNet. Meanwhile, because that our method injected more general semantic feature, these methods can perform also well. And owing to the generality of the extracted features based on foundation model, the performances sometimes are lower than the specific-designed convolutional networks, such as the SegNeXt model on UMF-CMGR fusion results.

The fusion images with semantic information can help improve semantic segmentation. Here, some qualitative results are provide performed on the fusion images, as shown in Fig.10. The segmentation results of the first two rows illustrate that the jointly optimization proposed in Alg.1 is feasible. While affected by the labelling quality, the

TABLE 3. The comparision of state-of-art methods and the proposed method on RoadScene and TNO.

Method	RoadScene			TNO		
	SCD	EN	VIF	SCD	EN	VIF
FusionGAN [4]	1.319	6.506	0.422	0.779	6.894	0.393
DIDFusion [37]	1.613	6.672	0.585	1.253	6.905	0.615
RFN-Nest [38]	1.717	6.888	0.535	1.624	7.240	0.551
SeAFusion [5]	1.701	7.087	0.667	1.469	7.238	0.637
UMF-CMGR [31]	1.594	6.654	0.587	1.333	7.021	0.647
TarDAL [23]	1.635	7.156	0.609	1.381	7.361	0.603
SwinFusion [8]	1.682	6.985	0.744	1.528	7.135	0.657
U2Fusion [17]	1.723	6.946	0.586	1.531	7.178	0.613
PSFusion [1]	1.791	7.413	0.734	1.704	7.509	0.664
Ours	1.821	7.490	0.762	1.711	7.486	0.673

TABLE 4. Quantitative results of different segmentation methods on MFNet dataset.

	Method	Infrared	Visible	UMF-CMGR	SwinFusion	SeAFusion	PSFusion	Ours
Person	BANet [32]	70.46	59.94	72.20	72.24	74.56	76.81	76.87
	SegFormer [33]	72.28	65.54	74.54	75.11	75.14	76.43	76.53
	SegNeXt [34]	72.47	65.76	75.19	75.32	75.93	77.66	77.78
	DINOv2-Seg [13]	72.62	65.81	75.21	75.30	76.03	77.79	77.86
Car Stop	BANet	65.85	71.43	70.53	70.85	74.38	74.32	74.42
	SegFormer	70.02	77.49	77.82	77.56	79.12	80.81	80.86
	SegNeXt	74.93	78.24	79.30	78.23	79.91	80.15	80.22
	DINOv2-Seg	74.89	78.34	78.42	78.19	79.96	80.24	80.29
Bike	BANet	69.23	70.00	71.20	70.31	72.09	72.93	72.78
	SegFormer	70.28	71.48	72.44	72.30	72.47	73.07	73.42
	SegNeXt	70.75	72.33	73.16	72.95	72.12	73.64	73.56
	DINOv2-Seg	70.78	72.42	73.30	72.43	72.52	73.59	73.65
Bump	BANet	72.72	75.31	75.74	80.36	78.35	81.01	79.97
	SegFormer	74.70	78.38	76.69	81.12	80.54	80.28	81.02
	SegNeXt	74.26	80.00	78.38	76.72	81.39	79.58	80.23
	DINOv2-Seg	74.46	80.06	77.82	81.16	81.43	80.54	81.13

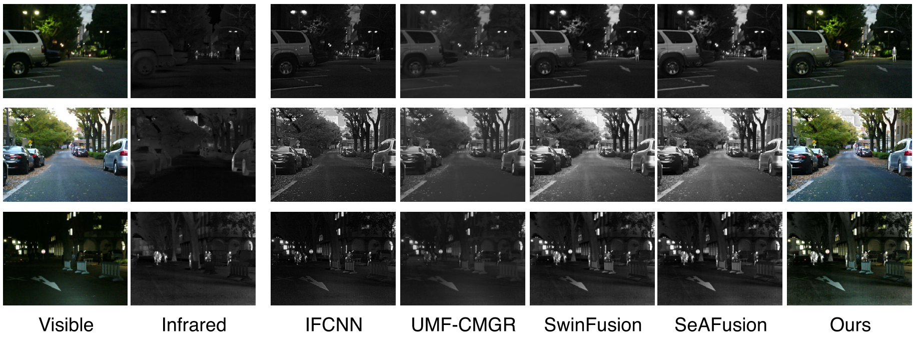

FIGURE 8. Qualitative results of different fusion methods on MFNet, and each row represents a different image pair.

segmentation result of the human in first row is inferior to the one in Fig.7, but with better boundaries. Compared with SeAFusion, which also can perform semantic segmentation, our method achieves better results, e.g., the people in remote distance can be accurately segmented as shown in the third row.

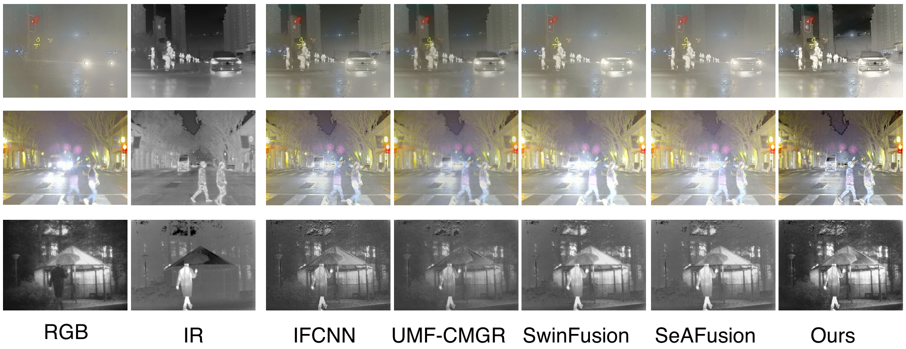

FIGURE 9. Qualitative results on M^3FD (the first row), RoadScene (the second row) and TNO (the third row).

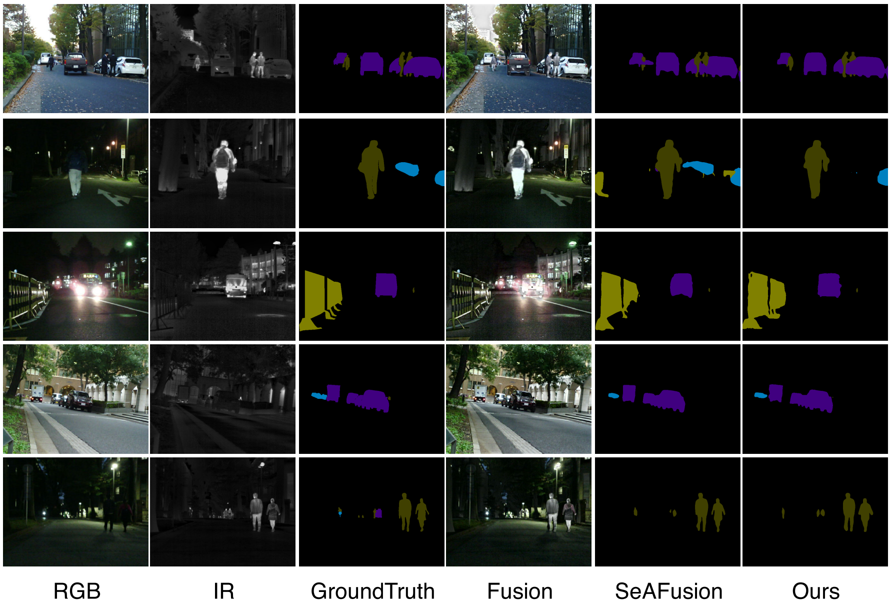

FIGURE 10. Qualitative segmentation results of the fusion images on MFNet, and each row denotes a different image pair.

V. CONCLUSION AND DISCUSSION

This paper resents a hybrid joint fusion and segmentation learning framework by introducing the general features from vision foundation models. Based on the hybrid learning idea, the fusion network can absorb the semantic features from CLIP with the DINOv2-based segmenter as a constraint. Meanwhile, with the feature injection network, the features between different-level tasks can be align to improve their performances. And further, benefit from the generalization ability of vision foundation models, the proposed network can be optimized mutually to promote the training process. Both quantitative and qualitative results on four datasets demonstrate the comparable performance with state-of-art methods. While limited by the object detection performance of DINOv2-based model, our method cannot carry out the object detection and segmentation simultaneously. In the further work, multi-task learning idea can be used, such as MaskDINO [35], to unify these tasks in one framework and lighten the model to speed up the inference stage. And further, more complicate strategies in image enhancement and super-resolution [25] can be used in feature fusion module to promote the performance.

REFERENCES

[1] L. Tang, H. Zhang, H. Xu, and J. Ma, "Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101870.

[2] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[3] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, doi: 10.1109/TIM.2021.3075747.

[4] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019, doi: 10.1016/j.inffus.2018.09.004.

[5] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022, doi: 10.1016/j.inffus.2021.12.004.

[6] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "DetFusion: A detection-driven infrared and visible image fusion network," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4003–4011.

[7] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13955–13965.

[8] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[9] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.

[10] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021, doi: 10.1109/TIM.2020.3038013.

[11] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, doi: 10.1109/TIM.2021.3056645.

[12] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[13] M. Oquab, "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[14] D. Jiang, "From CLIP to DINO: Visual encoders shout in multi-modal large language models," 2023, *arXiv:2312.0231*.

[15] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020, doi: 10.1109/TIP.2020.2975984.

[16] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "CGTF: Convolution-guided transformer for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[17] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022, doi: 10.1109/TPAMI.2020.3012548.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 15979–15988, doi: 10.1109/CVPR52688.2022.01553.

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

[20] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[21] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving CNN efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5977–5986, doi: 10.1109/CVPR.2017.633.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[23] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5792–5801.

[24] 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, doi: 10.1109/IROS.2017.8206396.

[25] X. Cao, Y. Lian, J. Li, K. Wang, and C. Ma, "Unsupervised multi-level spatio-spectral fusion transformer for hyperspectral image super-resolution," *Opt. Laser Technol.*, vol. 176, Sep. 2024, Art. no. 111032.

[26] A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, Dec. 2017, doi: 10.1016/j.dib.2017.09.038.

[27] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522, doi: 10.1117/1.2945910.

[28] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, Dec. 2015, doi: 10.1016/j.aeue.2015.09.004.

[29] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013, doi: 10.1016/j.inffus.2011.08.002.

[30] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020, doi: 10.1016/j.inffus.2019.07.011.

[31] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 3508–3515.

[32] Y. Chen, G. Lin, S. Li, O. Bourahla, Y. Wu, F. Wang, J. Feng, M. Xu, and X. Li, "BANet: Bidirectional aggregation network with occlusion handling for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3792–3801.

[33] E. Xie, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[34] M.-H. Guo, C.-Z. Lu, Q. Hou, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1140–1156.

[35] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3041–3050.

[36] S. Kan, Z. He, Y. Cen, Y. Li, V. Mladenovic, and Z. He, "Contrastive Bayesian analysis for deep metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7220–7238, Jun. 2023.

[37] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, and P. Li, "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Yokohama, Japan, Jul. 2020, pp. 970–976.

[38] H. Li, X.-J. Wu, and J. Kittler, "RFN-nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.

YICHUANG LUO received the M.S. degree in control theory and engineering from Xi'an Technological University, Xi'an, in 2013.

From 2013 to 2022, he was an Algorithm Engineer at Merit Data Technology Company Ltd. Since 2023, he has been an Assistant Engineer with the Department of Intelligent Science and Information Engineering, Xi'an Peihua University, Shaanxi. His research interests include distributed algorithm and systems, object detection, and tracking, edge computing, and applications.

FANG WANG received the M.S. degree in biomedical engineering from Xi'an Technological University, Xi'an, in 2013.

From 2013 to 2021, she was a Hardware Engineer at Kunlan Technology Company Ltd. Since 2021, she has been a Lecturer with the Department of Intelligent Science and Information Engineering, Xi'an Peihua University, Shaanxi. Her research interests include intelligent robot and intelligent information processing and applications.

XIAOHU LIU received the M.S. and Ph.D. degrees in mechatronic engineering from Xi'an Technological University, Xi'an, in 2013 and 2023, respectively.

From 2021 to 2023, he was an Associate Professor with the Intelligent Science and Information Engineering Department, Xi'an Peihua University, Shaanxi. Since 2023, he has been an Associate Professor with the Trine Engineering Institute, Shaanxi University of Technology, Shaanxi. His research interests include intelligent information processing and application, information fusion and application, and object tracking.