CtrlFuse: Mask-Prompt Guided Controllable Infrared and Visible Image Fusion

Yiming Sun[1], **Yuan Ruan**[1], **Qinghua Hu**[2], **Pengfei Zhu**[1,3,4*]

[1]School of Automation, Southeast University, Nanjing, China
[2]School of Artificial Intelligence, Tianjin University, Tianjin, China
[3]Low-Altitude Intelligence Laboratory, Xiong'an National Innovation Center, Xiongan, China
[4]Xiong'an Guochuang Lantian Technology Co., Ltd., Xiongan, China
{sunyiming, ruanyuan}@seu.edu.cn, {huqinghua, zhupengfei}@tju.edu.cn

Abstract

Infrared and visible image fusion generates all-weather perception-capable images by combining complementary modalities, enhancing environmental awareness for intelligent unmanned systems. Existing methods either focus on pixel-level fusion while overlooking downstream task adaptability or implicitly learn rigid semantics through cascaded detection/segmentation models, unable to interactively address diverse semantic target perception needs. We propose CtrlFuse, a controllable image fusion framework that enables interactive dynamic fusion guided by mask prompts. The model integrates a multi-modal feature extractor, a reference prompt encoder (RPE), and a prompt-semantic fusion module (PSFM). The RPE dynamically encodes task-specific semantic prompts by fine-tuning pre-trained segmentation models with input mask guidance, while the PSFM explicitly injects these semantics into fusion features. Through synergistic optimization of parallel segmentation and fusion branches, our method achieves mutual enhancement between task performance and fusion quality. Experiments demonstrate state-of-the-art results in both fusion controllability and segmentation accuracy, with the adapted task branch even outperforming the original segmentation model.

Code — https://github.com/Sevryy/CtrlFuse

Introduction

Infrared and visible image fusion aims to combine complementary information from both modalities to generate comprehensive representations of scenes (Ma et al. 2016; Tang, Li, and Ma 2025), thereby enhancing the environmental perception capabilities of intelligent unmanned systems from night to day (Sun et al. 2024a, 2025). Although visible images provide rich color information and high spatial resolution, their performance in downstream tasks degrades under poor illumination conditions. Infrared images effectively compensate for the limitations of visible imaging in darkness through thermal target imaging, but lack texture information of targets, leading to misidentification issues in downstream applications. How to better leverage the advantages of both imaging modalities to improve various downstream applications has become a key research focus.

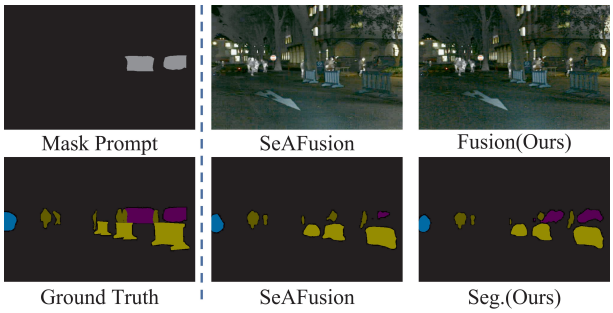

Figure 1: The importance of mask prompt-based interactive controllable image fusion for the performance of downstream application models (e.g., semantic segmentation).

With the development of deep learning, several representative infrared and visible image fusion methods have emerged (Cao et al. 2023). Autoencoder-based methods enhance fusion performance through multi-scale feature decomposition or pretraining-finetuning strategies. Generative adversarial network (GAN)-based (Ma et al. 2020) approaches preserve effective information from different modalities by establishing adversarial learning between fused images and multimodal source images. Beyond CNN-based methods, transformer architectures and diffusion models (Tang et al. 2024) have also attracted researchers' attention. Traditional image fusion tasks primarily focus on pixel-level consistency between fused and source images, evaluating performance through image quality assessment metrics, while neglecting the crucial requirement that fused images should effectively serve downstream perception tasks for improved task performance.

In recent years, researchers have recognized this need, leading to the proposal of task-related multimodal image fusion methods (Tang et al. 2025b; Liu et al. 2024a). These approaches connect image fusion networks with downstream task networks (e.g., object detection or semantic segmentation models) through joint optimization, enabling fusion models to implicitly learn semantically relevant information for downstream tasks. However, this paradigm severely restricts the fusion model's semantic perception capability to the predefined recognition types of the cascaded down-

*Corresponding author

stream task model, failing to dynamically control the attention to specific targets according to varying application requirements. As shown in Fig. 1, although existing methods have learned target semantics during training, they still struggle to adapt to real-world vehicle segmentation scenarios. Our method significantly enhances the semantic segmentation capability for designated targets (Car) through controllable prompt guidance, improving the practicality of task-driven image fusion models. Therefore, constructing semantically controllable multimodal image fusion architectures that enable dynamic controllable fusion according to different semantic requirements could bridge high-level downstream applications with low-level image fusion tasks.

In this paper, we propose a controllable image fusion method with multimodal semantic-aware prompt tuning (CtrlFuse), establishing a semantic-guided controllable multimodal image fusion framework through prompt tuning of the foundation model with superior semantic perception capability (Segment Anything Model (Kirillov et al. 2023), SAM). Specifically, the proposed CtrlFuse consists of four components: a multimodal backbone encoder-decoder, a reference prompt encoder (RPE), a prompt semantic fusion module (PSFM), and the pre-trained SAM. In the proposed RPE, we construct support features and query features that form prompt features for SAM fine-tuning under mask guidance. The PSFM explicitly fuses semantic prompt features, segmentation masks from SAM, and multimodal features from the encoder, achieving dynamic explicit injection of semantic information. Our method enhances performance on multimodal image fusion tasks by jointly optimizing the loss functions for both the SAM task branch and the image fusion branch. The main contributions are summarized as follows:

- We propose an interactively controllable multimodal image fusion framework that establishes a concise controllable fusion paradigm through dynamic mask prompts and explicit fusion of multimodal features with foundation model fine-tuning.
- We design a reference prompt encoder to dynamically generate semantic prompts for SAM fine-tuning, and develop a prompt semantic fusion module to explicitly aggregate semantic prompts with multimodal features, enhancing semantic perception capability.
- Extensive experiments validate the superior fusion performance and semantic controllability of our method. In particular, our approach achieves enhanced fusion capability through multimodal semantic fine-tuning, revealing the synergistic advantage of mutual promotion between the fusion and segmentation tasks.

Related Works

Infrared and Visible Image Fusion integrates complementary multi-modal features to preserve salient information from both sources (Ma, Ma, and Li 2019; Zhang et al. 2021; Liu et al. 2025). Ma et al. (Ma et al. 2019) proposed FusionGAN to formulate the image fusion task as an adversarial game between preserving infrared thermal radiation and visible texture details. Zhao et al. (Zhao et al. 2020) proposed a deep image decomposition framework that separates source images into background and detail feature maps, followed by dedicated fusion strategies for each feature. Li et al. (Li, Wu, and Durrani 2020) proposed a fusion network featuring nested connectivity to achieve multi-scale fusion of multimodal images. Recently, some task-driven methods have also attracted wide attention. Sun et al. (Sun et al. 2022a) proposed a detection-driven network for infrared and visible image fusion that leverages target-specific features from object detection tasks to guide the fusion process. Liu et al. (Liu et al. 2022) proposed a fusion framework that enhances object detection performance through dual adversarial optimization. Tang et al. (Tang, Yuan, and Ma 2022) proposed a semantic-aware fusion network that cascades the fusion module with a semantic segmentation module to enhance fusion performance. Yi et al. (Yi et al. 2024) proposed a text-guided image fusion framework that utilizes prompts to enable degradation-aware fusion. However, existing task-driven methods typically cascade image fusion with downstream application tasks, with high-level visual models (detection or segmentation) guiding the optimization of fusion models. This approach only supports the learning of fixed semantic categories and is difficult to adapt dynamically to complex and diverse semantic category perception requirements. Furthermore, it cannot achieve interactive and controllable fusion.

Interactive Deep Learning Model refers to a class of neural network architectures that enable human-steered regulation of model behavior (Cao et al. 2024; Tang et al. 2025a). Kirillov et al. (Kirillov et al. 2023) proposed the Segment Anything Model (SAM), a general-purpose segmentation system featuring interactive control and strong zero-shot generalization. Based on SAM, many other interactive large models have also been proposed, such as VRP-SAM (Sun et al. 2024b), GroundedSAM (Ren et al. 2024), and SAGE (Wu et al. 2025). In recent years, diffusion models have demonstrated significant progress in controllability (Shi et al. 2024; Zhang and Agrawala 2024; Hoe et al. 2024). Avrahami et al. (Avrahami, Lischinski, and Fried 2022) developed blended diffusion by effectively integrating CLIP's (Radford et al. 2021) semantic understanding capabilities with diffusion models to achieve text-driven image inpainting. Huang et al. (Huang et al. 2023) proposed a novel paradigm for composable conditional image synthesis that achieves creative controllability while maintaining generation quality. However, existing fusion methods predominantly focus on optimizing quantitative metrics for static outputs, while largely neglecting the development of controllable dynamics. This oversight results in limited adaptability to real-world scenarios requiring interactive adjustment or dynamic responses. Considering that image fusion should ultimately enhance downstream task performance, controllable region-aware fusion enhancement methods would provide greater practical value.

Methods

Overall Architecture

In this paper, we propose a semantic-aware framework enabling controllable fusion via multi-modal prompt tuning,

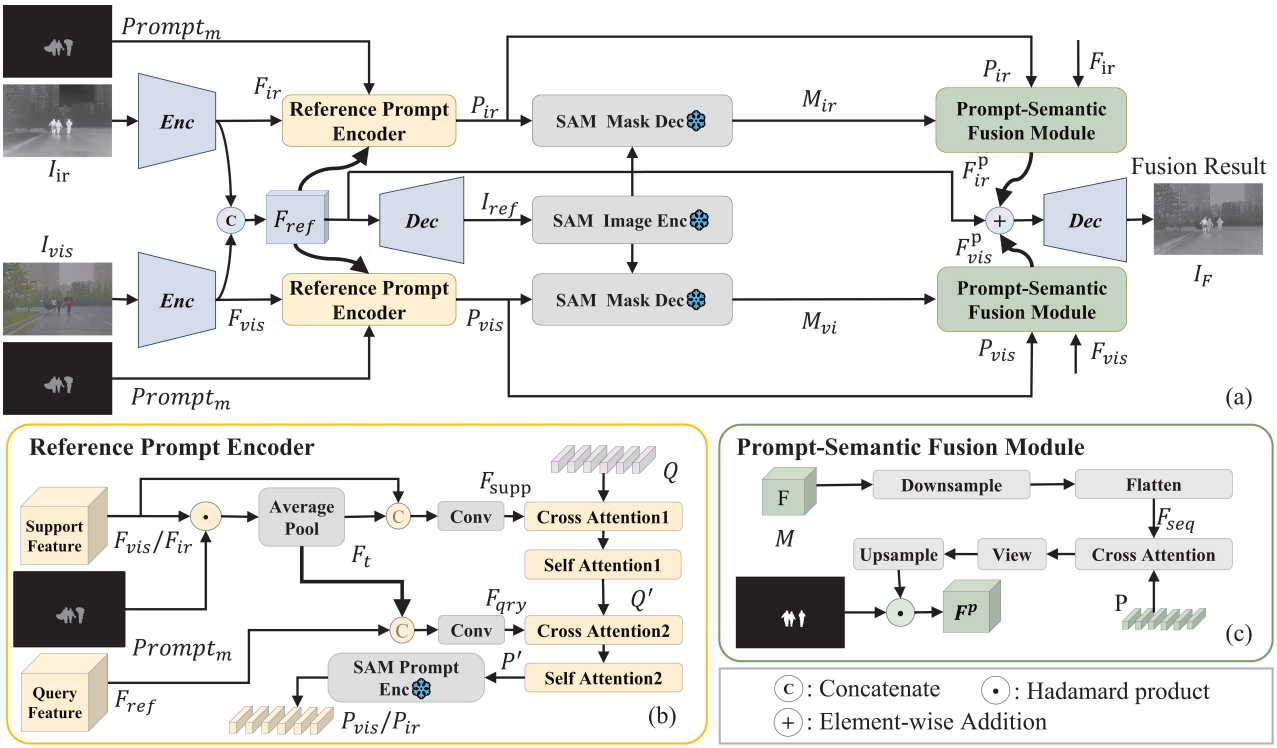

Figure 2: The architecture of CtrlFuse. CtrlFuse consists of two reference prompt encoders, two Prompt-Semantic fusion modules guided by the prompt mask, a set of infrared and visible feature encoders, and the auxiliary network.

termed CtrlFuse. In Fig 2, CtrlFuse contains a Reference Prompt Encoder, a Prompt-Semantic Fusion Module, and the auxiliary network.

In Fig. 2(a), we send a pair of infrared image $I_{ir} \in \mathbb{R}^{1\times H\times W}$ and visible image $I_{vis} \in \mathbb{R}^{3\times H\times W}$ into the infrared and visible encoders to extract the features F_{ir} and F_{vis}, respectively. The structure of encoders follows (Tang, Yuan, and Ma 2022). Then we concatenate F_{ir} and F_{vis} into F_{ref} as the input of the image decoder consisting of convolutional layers and activation layers to get the reference image I_{ref}. We send a mask as a prompt into Reference Prompt Encoder along with F_{ir} or F_{vis} and F_{ref}. In the Reference Prompt Encoder, we use F_{ir} or F_{vis} as support feature, F_{ref} as query feature and get prompt feature embedding P_{ir} and P_{vis}, respectively. We then send I_{ref} into the frozen SAM image encoder, and the result is sent to frozen SAM mask decoder with P_{ir} and P_{vis} respectively, after that we get two prediction masks M_{ir} and M_{vis}. To obtain the dual-light prompt features F_{ir}^{p} and F_{vis}^{p}, we input the prediction masks, prompt embeddings, and encoded features into the Prompt-Semantic Fusion Module. Through element-wise addition of the preliminary fusion feature F_{ref} with the prompt features F_{ir}^{p} and F_{vis}^{p}, the final fusion features are obtained. These final features are subsequently input into the image decoder to generate the ultimate fused image $I_{\mathcal{F}} \in \mathbb{R}^{1\times H\times W}$.

We train the proposed CtrlFuse model in an end-to-end manner. The model is optimized mainly by calculating both the fusion loss $\mathcal{L}_{\text{fusion}}$ and the segmentation loss $\mathcal{L}_{\text{seg}}$.

Reference Prompt Encoder

As shown in Fig. 2(b), we propose a Reference Prompt Encoder to dynamically extract prompt embeddings from the controllable regions of interest.

Taking the infrared reference prompt encoder as an example, we take F_{ir} as a support feature, F_{ref} as a query feature, and input them along with $Prompt_m$ into the encoder. To enhance the features in the regions of interest, we compute the Hadamard product between $Prompt_m$ and F_{ir}, followed by an average pooling operation to get F_t. This process can be described as follows:

$$F_{\mathrm{t}} = \mathrm{AveragePool}\left(\mathrm{Prompt}_{\mathrm{m}} \cdot F_{\mathrm{ir}}\right). \tag{1}$$

In order to achieve a holistic feature representation that captures both detailed local information and global context, which enhances the model's performance and generalization ability, we separately concatenate F_{ir} and F_{ref} with F_t, and then apply convolutional layers to produce the resulting features $F_{\mathrm{supp}} \in \mathbb{R}^{C\times H\times W}$ and$F_{\mathrm{qry}} \in \mathbb{R}^{C\times H\times W}$. This process can be described as follows:

$$F_{supp} = \mathrm{Conv}(\mathrm{Concat}(F_{\mathrm{ir}}, F_{\mathrm{t}})), \tag{2}$$

$$F_{qry} = \mathrm{Conv}(\mathrm{Concat}(F_{\mathrm{ref}}, F_{\mathrm{t}})). \tag{3}$$

We get a set of learnable queries $Q \in \mathbf{R}^{N\times C}$, that extract different aspects of information from F_{supp}, where N corresponding to the number of prominent features the model focuses on, is set to 40 as determined by hyperparameter experiments. We first pass the queries Q and F_{supp}

through a cross-attention layer, and then feed the resulting output into a self-attention layer to obtain queries that are controllable for specific categories in the support image $Q^{'}$. Subsequently, $Q^{'}$ and F_{qry} are processed through a cross-attention mechanism. The resulting output is then fed into a self-attention layer to derive a set of reference prompts $P^{'}$, which correspond to specific categories present in the query image. This process can be described as follows:

$$Q' = \mathrm{SelfAttn}_1\left(\mathrm{CrossAttn}_1\left(Q, F_{supp}\right)\right) \tag{4}$$

$$P' = \mathrm{SelfAttn}_2\left(\mathrm{CrossAttn}_2\left(Q', F_{qry}\right)\right) \tag{5}$$

Then, $P^{'}$ is passed through the frozen SAM Prompt Encoder to generate the final prompt feature embedding P.

Prompt-Semantic Fusion Module

As shown in Fig. 2(c), we propose a Prompt-Semantic Fusion Module to obtain category-specific prompt features F_{ir}^{p} and F_{vis}^{p}. In both the infrared and visible branches, the initial encoded features F, the prompt feature embeddings P, and the corresponding SAM segmentation outputs M are used as input representations. Given the notable differences in how various modalities contribute to semantic information, we employ segmentation results derived from each modality-specific branch (e.g., infrared and visible) as masks. Specifically, the quality of the segmentation outcomes directly reflects the richness and quality of the high-level features provided by that modality, thereby indicating its potential contribution to overall task performance. Such a strategy not only facilitates the quantification of each modality's effectiveness but also provides more precise guidance for multi-modal data fusion. First, we downsample the feature map $F \in \mathbb{R}^{H\times W\times C}$ to reduce the computational load of the network. Then, we flatten the downsampled features into a sequence format $F_{seq} \in \mathbf{R}^{(H*W)\times C}$. This process can be described as follows:

$$F_{seq} = Flatten(Down(F)) \tag{6}$$

We process the serialized features F_{seq} together with the previously derived prompt embeddings P through a cross-attention mechanism. Subsequently, the attended features are subjected to a view transformation operation to revert to their original spatial dimensions, followed by an upsampling step to restore them to the original resolution, thereby obtaining enhanced spatial features. Finally, these enhanced spatial features are element-wise multiplied with the corresponding segmentation masks to derive the enhanced features specific to the indicated categories . This process can be described as follows:

$$F^p = M \cdot (Up(View(CrossAttn(F_{seq}, P)))) \tag{7}$$

Experiments

Experimental Setting

Implementation Details. We performed experiments on a computing platform with four NVIDIA GeForce RTX 3090 GPUs. We used Adam Optimization to update the overall network parameters with the learning rate of 1.0×10^{-4}. The training epoch is set to 150 and the batch size is 4.

Datasets and Partition Protocol. We conducted experiments on three publicly available datasets: FMB (Liu et al. 2023), MSRS (Tang et al. 2022), and DroneVehicle (Sun et al. 2022b). FMB and MSRS contain labels for semantic segmentation, and we generate masks for different categories from these labels. FMB contains $1,500$ infrared-visible image pairs captured by onboard cameras. We used $1,020$ pairs of images for training, 280 pairs of images for validation, and the remaining 200 pairs for evaluation. MSRS contains $1,444$ infrared-visible image pairs captured by onboard cameras. We used $1,072$ image pairs for training, 150 image pairs for validation, and 200 image pairs for evaluation. For DroneVehicle, we generate masks using other semantic segmentation models. We evaluated 200 image pairs from DroneVehicle. We present both qualitative and quantitative analyses of the MSRS dataset in the supplementary material.

Competing Methods. We compared 8 state-of-the-art methods on three publicly available datasets. In these comparison methods, LDfusion (Wang et al. 2024) is the CLIP-based image fusion method, NestFuse (Li, Wu, and Durrani 2020) is the autoencoder-based method, DIDFuse (Zhao et al. 2020) and CDDFuse (Zhao et al. 2023) are the deep learning-based image decomposition methods. SwinFuse (Wang et al. 2022) is a Transformer-based method. SeAFusion (Liu et al. 2022) and SDCFusion (Liu et al. 2024b) are the segmentation-driven methods. PSFusion (Tang et al. 2023) is a fusion method driven by high-level vision tasks.

Evaluation Metrics. We evaluated the performance of the proposed method based on qualitative and quantitative results. The qualitative evaluation is mainly based on the visual effect of the fused image. A good fused image needs to have complementary information from multi-modal images. The quantitative evaluation mainly uses quality evaluation metrics to measure the performance of image fusion. We selected 6 popular metrics, including the MSE, PSNR, N_{abf} , gradient-based similarity measurement (Q_{abf}) (Xydeas and Petrovic 2000),SSIM (Wang et al. 2004) and SCD. We also evaluate the performance of the different methods on the typical downstream task, infrared-visible object detection, and semantic segmentation.

Evaluation on the FMB Dataset

Quantitative Comparisons. The quantitative results on the FMB dataset are summarized in Table 1. Our method outperforms 8 state-of-the-art approaches on three key metrics. Achieving the best scores in PSNR and N_{abf} on both the MSRS and FMB datasets demonstrates its consistent ability to preserve image clarity and reduce distortion across diverse scenarios. The top performance in Q_{abf} further indicates that our method effectively retains and integrates structural and textural information from the source images. A high N_{abf} score highlights strong gradient consistency and spatial coherence, which are crucial for vision-based tasks. These advantages make our method well-suited for downstream applications such as object detection.

Qualitative Comparisons. We mark the foreground region with the yellow rectangular box showing their zoomed-in effects for easier comparison in Fig. 3. Among all methods ex-

Methods	MSE	PSNR	Q_{abf}	N_{abf}	SSIM	SCD
FMB Dataset						
LDFusion	0.061	60.71	0.51	0.112	0.514	1.549
SwinFuse	**0.042**	62.334	0.577	0.029	0.905	1.9
NestFuse	0.046	61.96	0.483	0.042	0.787	1.594
CDDFuse	0.048	**62.696**	0.674	**0.026**	**1.002**	**1.626**
DIDFuse	0.047	61.565	0.528	0.042	0.765	**1.824**
SeAFusion	0.047	62.539	0.654	0.029	**0.964**	1.62
PSFusion	0.051	61.517	0.627	0.056	0.836	**1.875**
SDCFusion	0.048	62.456	**0.693**	0.031	0.906	1.657
CtrlFuse(Ours)	**0.043**	**63.292**	**0.719**	**0.024**	0.925	1.522
DroneVehicle Dataset						
LDFusion	0.076	59.573	0.376	0.054	0.568	1.38
SwinFuse	0.084	59.165	0.202	0.069	0.558	1.295
NestFuse	0.071	59.786	0.307	0.052	0.486	1.413
CDDFuse	**0.065**	**60.199**	0.469	**0.021**	0.845	1.359
DIDFuse	0.067	59.988	0.265	0.062	0.466	1.459
SeAFusion	0.094	58.649	0.492	0.044	**0.879**	1.472
PSFusion	0.067	60.065	0.454	0.095	0.717	**1.534**
SDCFusion	0.078	59.443	**0.534**	**0.035**	**0.853**	1.316
CtrlFuse(Ours)	**0.063**	**60.317**	**0.496**	**0.035**	0.779	**1.552**
MSRS Dataset						
LDFusion	0.056	61.05	0.438	0.116	0.541	1.515
SwinFuse	0.038	63.69	0.178	0.026	0.343	1.033
NestFuse	**0.033**	64.128	0.242	0.025	0.217	1.138
CDDFuse	0.038	64.309	**0.689**	0.023	**1.001**	1.623
DIDFuse	**0.035**	63.94	0.204	0.025	0.223	1.121
SeAFusion	0.036	**64.491**	0.675	**0.021**	**0.982**	1.707
PSFusion	0.037	64.001	0.676	0.042	0.917	**1.812**
SDCFusion	0.039	64.003	**0.712**	0.023	0.957	**1.739**
CtrlFuse(Ours)	**0.035**	**64.75**	0.685	**0.018**	0.969	1.726

Table 1: Quantitative comparison of CtrlFuse with 8 state-of-the-art methods. **Bold underline** indicates the best, **Bold** indicates the second best, and Underlined indicates the third.

cept LDFusion and our proposed method, the fused images suffer from excessive brightness information in the truck's windshield due to over-enhanced infrared information, resulting in blurry appearances. Only in the fused images produced by LDFusion and our method can the person inside the truck be clearly observed. Furthermore, our method better highlights the person, making them more distinguishable from the background.

Evaluation on the DroneVehicle Dataset

Quantitative Comparisons. Table 1 reports the performance of the different methods on the DroneVehicle dataset for 6 metrics, where our method achieves the best in 3 metrics . Among them, PSNR and MSE indicate that our method introduces minimal distortion and produces clearer and more detailed images with less noise interference. Moreover, the highest SCD indicates the method's effectiveness in maintaining sharpness and clarity. These quantitative results indicate that the proposed CtrlFuse method can efficiently capture and integrate multi-modal information, leading to fusion outcomes that are both visually compelling and rich in detail.

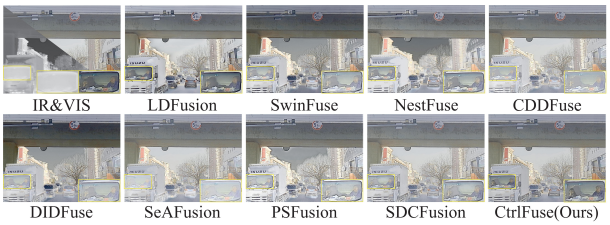

Figure 3: Qualitative comparisons of various methods on representative images selected from the FMB dataset.

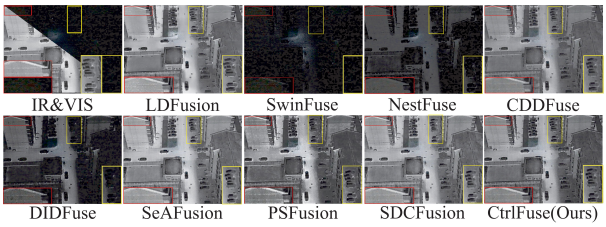

Figure 4: Qualitative comparisons of various methods on representative images selected from DroneVehicle dataset.

Qualitative Comparisons. We mark the background regions with yellow and red boxes, respectively, in Fig. 4, with zoomed-in views provided for clearer comparison.As indicated by the yellow box, in the fused images of SwinFuse, NestFuse, and DIDFuse, excessive preference for the visible image makes the scene overly dark, rendering the cars in shadowed areas indistinguishable. As highlighted by the red box, LDFusion fails to adequately preserve texture details, causing the stripes on the exterior wall to be smoothed out, while SwinFuse and NestFuse still fail to distinguish the details due to insufficient infrared information. Our fused image achieves a superior balance between the infrared and visible input images.

High-level Vision Tasks Evaluation

Infrared and visible image fusion integrates information from different spectral bands to produce more informative and comprehensive representations, commonly used in high-level vision tasks like object detection, classification, and scene understanding. In this section, we conduct experiments on semantic segmentation and object detection.

Segmentation Performance. We conducted quantitative experiments on the MSRS dataset. Please refer to the supplementary material for detailed experimental settings. Segmentation performance, reported in Table 2, is evaluated using pixel-wise IoU. As shown, our method achieves the best performance on four categories and ranks second on four others, with the highest overall mIoU, indicating superior segmentation accuracy and generalization. We also provide visual results in Fig. 5. For the "car" class (purple), only our method achieved complete segmentation,

MSRS	Background	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Tone	Bump	mIoU
Infrared	0.9669	0.7542	0.5251	0.7287	0.5926	0.7381	0.9763	0.7925	0.9299	0.7783
Visible	0.9658	0.7612	0.3904	0.7393	0.5664	0.7957	0.9797	0.7964	0.9414	0.7707
LDFusion	0.9667	0.6859	0.4463	0.7221	0.5078	0.7783	0.9837	0.7867	0.9404	0.7576
SwinFuse	0.9572	0.7013	0.4243	0.7090	0.5738	0.7757	0.9808	0.7561	0.9320	0.7567
NestFuse	0.9587	0.6844	0.4136	0.6938	0.5731	0.7758	0.9714	0.7434	0.9349	0.7499
CDDFuse	0.9696	0.7609	0.4856	0.7424	0.5843	0.7833	0.9829	0.7951	0.9375	0.7824
DIDFuse	0.9666	0.7471	0.4625	0.7268	0.5651	0.8009	0.9817	0.7807	0.9334	0.7739
SeAFusion	0.9695	0.7667	0.4886	**0.7507**	0.5991	0.7771	0.9822	0.7949	0.9395	0.7854
PSFusion	0.9708	0.7722	**0.5276**	0.7482	0.6122	0.7962	0.9812	0.8064	**0.9443**	0.7955
SDCFusion	**0.9708**	0.7682	0.5227	0.7420	0.6112	**0.8028**	0.9809	**0.8121**	0.9441	0.7950
CtrlFuse(Ours)	0.9705	**0.7810**	0.5134	0.7443	**0.6179**	0.8025	**0.9851**	0.8082	0.9438	**0.7963**

Table 2: Segmentation performance (mIoU) of visible, infrared and fused images on the MSRS dataset. **Bold** indicates the best, Underlined indicates the second best.

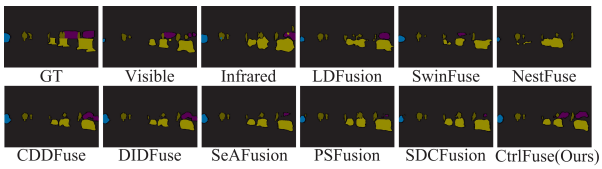

Figure 5: Segmentation results for infrared, visible, and fused images from the MSRS dataset. The segmentation models are retrained.

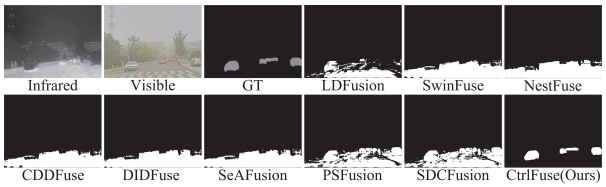

Figure 6: Segmentation results of the original SAM model and CtrlFuse on the FMB dataset.

further demonstrating its effectiveness in preserving critical semantic information for downstream tasks. We also evaluated DeepLabV3+ (Chen et al. 2018), pre-trained on Cityscapes (Cordts et al. 2016), on the FMB dataset, including the original dual-modal images, results from 8 SOTA methods, and our fused images. More details are provided in the supplementary material.

Detection Performance. Object detection is a fundamental and high-level task in computer vision that aims to identify and localize multiple objects within an image or video. To adapt the model to the DroneVehicle dataset, we selected the first 2000 images from the test set. Of these, 200 were randomly chosen for evaluation, and the remaining 1800 were used to fine-tune YOLOv5 on infrared-visible image pairs, improving its detection performance on this dataset.The fine-tuned YOLOv5 model was then used to evaluate fusion methods for object detection performance. The evaluation results are presented in Table 3, which shows the average performance over thresholds from 0.5 to 0.95

Method	car	truck	bus	freight car	All
LDFusion	0.632	0.409	0.475	0.324	0.460
SwinFuse	0.382	0.032	0.426	0.19	0.258
NestFuse	0.555	0.271	0.365	0.317	0.377
CDDFuse	0.632	0.455	0.470	0.387	0.486
DIDFuse	0.528	0.279	0.388	0.264	0.364
SeAFusion	0.646	0.458	0.514	**0.415**	0.508
PSFusion	0.628	0.475	0.424	0.277	0.451
SDCFusion	0.647	**0.526**	0.468	0.397	0.510
CtrlFuse(Ours)	**0.651**	0.520	**0.521**	0.409	**0.525**

Table 3: Object detection performance (AP@[0.5:0.95]) on the DroneVehicle dataset. **Bold** indicates the best, Underlined indicates the second best.

(AP@[0.5:0.95]). As can be observed, the proposed method achieves the best performance in the "car" and "bus" categories, ranks second in the remaining categories, and maintains the highest overall metric (All), demonstrating its superior object detection accuracy across multiple evaluation criteria. In addition, we provide visual comparison results in the supplementary materials. We also provide a series of comparative experiments on downstream tasks without prompt masks in the supplementary materials.

Ablation Study

We conducted ablation studies on the MSRS dataset and reported the results in Table 4.

w/o Prompt. To investigate the impact of prompts on fusion quality, we remove the prompt mask from our framework, retaining only the two image encoders and a single decoder. The results in Table 4 show that, while the pixel-level fidelity remains similar, the perceptual quality and structural integrity are notably improved. This indicates that the prompt mask enhances the suitability of fused images for high-level vision tasks, with improvements in structural similarity, noise reduction, and spectral consistency likely benefiting downstream tasks such as object detection and semantic segmentation.

Ablation	MSE	PSNR	Q_{abf}	N_{abf}	SSIM	SCD
w/o Prompt	0.035	64.958	0.637	**0.014**	0.933	1.635
w/o Seg	0.036	64.615	0.671	0.021	0.939	1.636
w/o Vis	**0.033**	**65.064**	0.656	0.02	0.915	1.681
w/o Ir	0.034	65.027	0.67	0.019	0.938	1.622
Exchange SQ	0.034	64.917	0.661	0.021	0.924	1.659
CtrlFuse(Ours)	0.035	64.75	**0.685**	0.018	**0.969**	**1.726**

Table 4: Ablation study on the MSRS dataset. **Bold** indicates the best, Underlined indicates the second best.

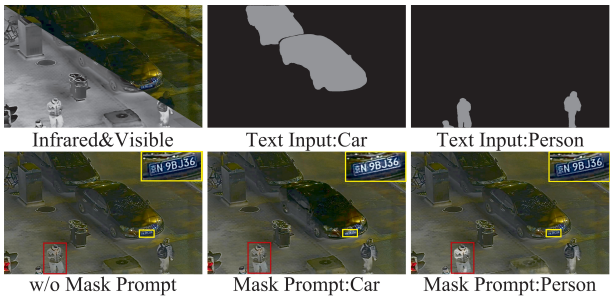

Figure 7: Fusion results of CtrlFuse with different prompt masks generated by Grounded-SAM on the LLVIP dataset.

w/o Seg. To verify that segmentation does not harm fusion performance, we remove the SAM mask decoder and the corresponding segmentation loss $\mathcal{L}_{seg}$. The resulting performance drop across all metrics demonstrates that segmentation improves fusion quality by preserving details, reducing noise, and maintaining structural and spectral integrity. Thus, segmentation plays a beneficial and essential role in achieving high-quality image fusion.

w/o Vis. To verify the contribution of both modalities, we remove the reference prompt encoder and prompt-semantic fusion module from the visible branch, eliminating the semantic prompt F_{vis}^{p} in fusion. While pixel-level fidelity slightly improves, the removal degrades perceptual quality and structural integrity, indicating that the visible prompt branch plays a key role in enhancing fusion performance.

w/o Ir. We perform the same ablation on the infrared branch by removing its corresponding components. A similar degradation in perceptual quality and structural integrity was observed, confirming that both modalities are crucial for high-quality image fusion.

Exchange SQ. To determine whether using F_{ir} or F_{vis} as the support feature is more effective, we swap the roles of support and query features in the reference prompt encoder. In the modified version, F_{ref} serves as the support feature, while F_{ir} or F_{vis} becomes the query. The overall drop in performance indicates that the original design is more effective for feature alignment and fusion.

Analysis and Discussion

The Impact of Mask Prompt Fine-tuning on SAM Performance. To investigate how the fusion branch enhances segmentation via prompt fine-tuning, we compared CtrlFuse with the original SAM and other fusion methods. As shown in Fig. 6, our method, which directly combines the segmentation masks from its two branches, achieves superior performance, demonstrating that the fine-tuned CtrlFuse enables more effective personalized segmentation guided by mask prompts.

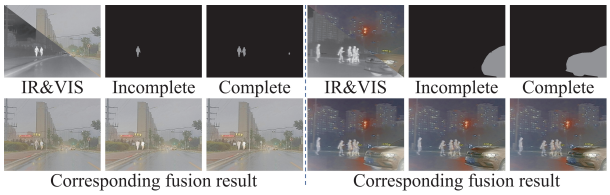

Figure 8: Comparison of fusion results under complete versus incomplete or coarse prompt masks.

Controllability of CtrlFuse. CtrlFuse uses prompt masks to guide the model to focus on and enhance specific targets. To validate this, we experimented on the LLVIP dataset (Jia et al. 2021), which lacks segmentation annotations. We employed Grounded-SAM (Ren et al. 2024) to generate masks from different text inputs and used them to guide the fusion process. The resulting images in Fig. 7 demonstrate the model's capability to emphasize designated classes under varying conditions.

Sensitivity Analysis on Prompt Mask. Our analysis shows the model's robustness to prompt mask quality. Even when masks are incomplete or of low quality (e.g., masking only one object or just a part of it), the fusion results still effectively highlight the target objects, including unmarked areas, as shown in Fig. 8. Visual comparisons confirm that mask quality has no significant impact on the output.

Conclusion

In this paper, we present CtrlFuse, a mask-prompt controllable multimodal image fusion framework that establishes dynamic interactions between scene understanding requirements and low-level fusion processes. By integrating mask-guided semantic adaptation with explicit feature fusion mechanisms, the proposed method addresses the critical limitation of conventional task-driven fusion approaches in rigid semantic constraints. Through synergistic design of the reference prompt encoder for adaptive semantic tuning, the prompt semantic fusion module for cross-modal feature aggregation, and the joint optimization of segmentation-aware fusion objectives, our framework achieves both semantically enhanced fusion results and improved downstream perception accuracy. Experimental validations across diverse scenarios demonstrate that the explicit injection of semantic guidance enables not only controllable fusion behavior but also mutual reinforcement between multimodal fusion and semantic segmentation tasks. The proposed paradigm provides an interactive and controllable multimodal fusion perception solution for intelligent unmanned systems, which is particularly suitable for all-weather search and rescue applications that require a focus on specific targets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62222608, 62506073, and 62436002, the Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270, the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20250395, the Jiangsu Funding Program for Excellent Postdoctoral Talent under Grant Number 2025ZB294, the the Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020004.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-Modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23555–23564.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2024. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*.

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

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Hoe, J. T.; Jiang, X.; Chan, C. S.; Tan, Y.-P.; and Hu, W. 2024. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6180–6189.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Jia, X.; Zhu, C.; Li, M.; Tang, W.; and Zhou, W. 2021. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3496–3504.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, H.; Wu, X.-J.; and Durrani, T. 2020. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12): 9645–9656.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 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 (CVPR)*, 5802–5811.

Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024a. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8115–8124.

Liu, J.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; Fan, X.; and Liu, R. 2025. Infrared and Visible Image Fusion: From Data Compatibility to Task Adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2349–2369.

Liu, X.; Huo, H.; Li, J.; Pang, S.; and Zheng, B. 2024b. A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion*, 108: 102352.

Ma, J.; Chen, C.; Li, C.; and Huang, J. 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31: 100–109.

Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45: 153–178.

Ma, J.; Xu, H.; Jiang, J.; Mei, X.; and Zhang, X.-P. 2020. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48: 11–26.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8839–8849.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022a. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4003–4011.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022b. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32: 6700–6713.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2024a. Dynamic brightness adaptation for robust multi-modal image fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 1317–1325.

Sun, Y.; Chen, J.; Zhang, S.; Zhang, X.; Chen, Q.; Zhang, G.; Ding, E.; Wang, J.; and Li, Z. 2024b. VRP-SAM: SAM with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23565–23574.

Sun, Y.; Li, X.; Zhu, P.; Hu, Q.; Ren, D.; Xu, H.; and Zhu, X. 2025. Task-Gated Multi-Expert Collaboration Network for Degraded Multi-Modal Image Fusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 57571–57586. PMLR.

Tang, L.; Deng, Y.; Yi, X.; Yan, Q.; Yuan, Y.; and Ma, J. 2024. DRMF: Degradation-Robust Multi-Modal Image Fusion via Composable Diffusion Prior. In *Proceedings of the ACM International Conference on Multimedia*, 8546–8555.

Tang, L.; Li, C.; and Ma, J. 2025. Mask-DiFuser: A Masked Diffusion Model for Unified Unsupervised Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Tang, L.; Wang, Y.; Cai, Z.; Jiang, J.; and Ma, J. 2025a. ControlFusion: A Controllable Image Fusion Framework with Language-Vision Degradation Prompts. *Advances in Neural Information Processing Systems*.

Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; and Ma, J. 2025b. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 133: 5262–5280.

Tang, L.; Yuan, J.; and Ma, J. 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: 28–42.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84: 79–92.

Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2023. 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. *Information Fusion*, 99: 101870.

Wang, Y.; Miao, L.; Zhou, Z.; Zhang, L.; and Qiao, Y. 2024. Infrared and visible image fusion with language-driven loss in CLIP embedding space. *arXiv preprint arXiv:2402.16267*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Chen, Y.; Shao, W.; Li, H.; and Zhang, L. 2022. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Transactions on Instrumentation and Measurement*, 1–1.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. *arXiv preprint arXiv:2503.01210*.

Xydeas, C. S.; and Petrovic, V. S. 2000. Objective image fusion performance measure. *Electronics Letters*, 36: 308–309.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27026–27035.

Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhang, L.; and Agrawala, M. 2024. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5906–5916.

Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Zhang, J.; and Li, P. 2020. DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 970–976. ijcai.org.