Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion

Zeyu Wang[1] Jizheng Zhang[2] Haiyu Song[1*] Mingyu Ge[1] Jiayu Wang[1] Haoran Duan[3*]
[1]College of Computer Science and Engineering, Dalian Minzu University
[2]Faculty of Science and Technology, University of Macau
[3]Department of Automation, Tsinghua University
{wangzeyu, shy}@dlnu.edu.cn, haoranduan@mail.tsinghua.edu.cn

Abstract

Infrared and visible image fusion (VIS-IR) aims to integrate complementary information from both source images to produce a fused image with enriched details. However, most existing fusion models lack controllability, making it difficult to customize the fused output according to user preferences. To address this challenge, we propose a novel weakly-supervised, instance-level controllable fusion model that adaptively highlights user-specified instances based on input text. Our model consists of two stages: pseudo-label generation and fusion network training. In the first stage, guided by observed multimodal manifold priors, we leverage text and manifold similarity as joint supervisory signals to train text-to-image response network (TIRN) in a weakly-supervised manner, enabling it to identify referenced semantic-level objects from instance segmentation outputs. To align text and image features in TIRN, we propose a multimodal feature alignment module (MFA), using manifold similarity to guide attention weight assignment for precise correspondence between image patches and text embeddings. Moreover, we employ spatial positional relationships to accurately select the referenced instances from multiple semantic-level objects. In the second stage, the fusion network takes source images and text as input, using the generated pseudo-labels for supervision to apply distinct fusion strategies for target and non-target regions. Experimental results show that our model achieves state-of-the-art fusion performance and accurately highlights user-defined instances. Code: https://github.com/GMY628/RIS-Fuse.

1. Introduction

Infrared-visible image fusion (VIS-IR) combines complementary features from infrared (IR) and visible (VIS) images to create more informative results [39, 40, 53]. IR images capture thermal radiation, enabling effective low-light and night-time detection, but lack texture details, while VIS images offer rich details but are limited by lighting conditions. Fusing these modalities aims to preserve thermal information from IR while enhancing it with the detail from VIS [28, 58]. This is crucial for applications like surveillance, target detection, and autonomous driving [1, 26, 44, 52].

Existing fusion methods face major limitations in controllability and user adaptability. Users, real-world scenarios, or diverse downstream tasks, frequently require varying levels of fusion and target emphasis. However, once trained, these static methods produce fixed outputs, failing to adapt to dynamic needs [4]. While recent works have attempted to introduce controllability into fusion through text-driven

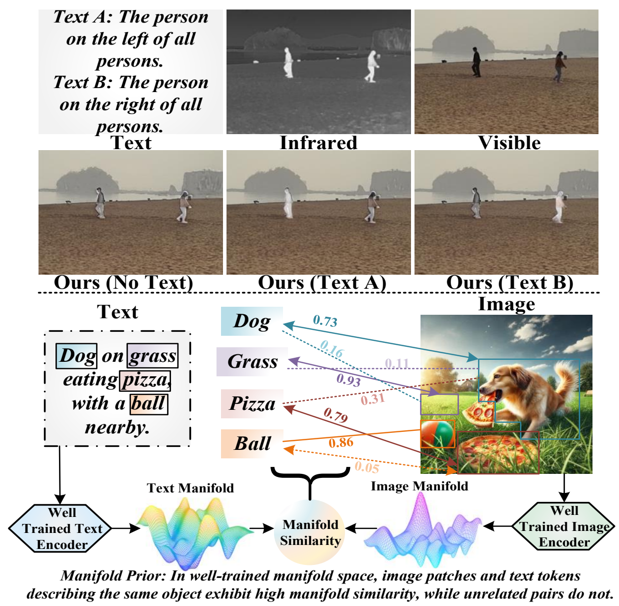

Figure 1. Top: Our model highlights instance objects guided by text. Bottom: Discovered manifold prior between text and image.

*Corresponding author

models [4, 19, 27, 42, 51, 54], these models are limited in scope, focusing on global adjustments or semantic-level highlighting, rather than enabling precise, instance-level control. Instance-level fusion control not only addresses dynamic user needs but also enhances downstream task performance, such as targeted object detection.

Achieving controllable image fusion (CIF) requires precise instance localization from text input. While existing image-text interaction paradigms enable instance localization in natural images [25, 30, 46], they are not directly applicable to VIS-IR images, as they are designed for single-modality natural images and overlook unique characteristics of infrared and visible modalities. For visible images, low-light conditions often obscure key objects, making localization unreliable. Infrared images, though effective in highlighting thermal information, lack texture, color, and fine-grained details [22], leading to two key challenges that hinder localization: (1) Non-thermal objects (e.g., background elements or low-temperature regions) lack distinctive features, making it harder to identify their categories. (2) Objects of the same category (e.g., people, vehicles) often share similar thermal signatures, lacking unique visual cues, which causes difficulty in instance differentiation. Despite the need, text-driven instance localization tailored for VIS-IR remains unexplored, limiting CIF advancements.

To address these challenges, we propose a weakly-supervised instance-level controllable fusion model that allows users to highlight the specified instance in fused images via text input, as shown in Fig. 1. Our model follows a two-stage training paradigm. Stage I employs weak supervision to generate pseudo-labels that capture instance localization from textual input, addressing the lack of instance-level annotations in existing VIS-IR datasets, without manual labeling; stage II trains the fusion network using these pseudo-labels, equipping it with both instance localization and image fusion capabilities. Stage I consists of a text-to-image response network (TIRN) and an instance selection module (ISM). TIRN identifies semantic-level objects based on user input and is trained in a weakly supervised manner using only manifold similarity (MS) and text as joint supervision signals [25, 62]. This is supported by an observed manifold prior: in well-trained feature spaces, for images and text represented as two separate manifolds [14, 36], image patches and text tokens describing the same object exhibit high MS, while unrelated pairs do not, as illustrated in Fig. 1. Moreover, inspired this prior, we tailor a multimodal feature interaction module (MFA) for VIS-IR-Text interactions in TIRN. MFA treats cross-modal misalignment as distortions in high-dimensional manifolds. By computing MS between image patches and text tokens, it leverages this similarity as a key factor for assigning attention weights, which significantly reduces the difficulty of instance object localization. After TIRN, ISM selects the referenced instance based on spatial relationships. Spatial cues from text provide an effective basis for localization by comparing the coordinates of the "centroid" pixels of multiple objects, without requiring training.

In stage II, the image fusion network is trained with pseudo-labels, text, and source images, learning implicit associations between pseudo-labels and text while selectively highlighting referenced objects. To balance object emphasis and overall image quality, we design separate fusion strategies for target and non-target regions.

The contributions are summarized as follows.

- We present the first instance-level controllable VIS-IR fusion model that localizes user-specified objects from text and highlights target regions with tailored fusion modes.
- We propose TIRN and ISM, the first text-driven weakly supervised instance localization method for VIS-IR. TIRN generates semantic-level localization maps using only text and MS as supervision, while ISM refines them via spatial constraints to create pseudo-labels for fusion training.
- We design MFA that uses MS between image patches and text tokens as a crucial factor for matching features, promoting alignment from a manifold perspective.
- We discover a key manifold prior that justifies using MS to supervise TIRN and align text-image features.

2. Related Work

2.1. Uncontrollable *vs.* Controllable Image Fusion

Uncontrollable image fusion (UIF) methods operate in a static manner, applying uniform fusion strategies across the entire image, regardless of specific task requirements or user preferences. While effective for general feature merging, UIF lacks flexibility and may dilute critical details, reducing effectiveness for target detection. Most fusion models, such as [15, 17, 60, 61, 63], fall into this category.

Controllable image fusion (CIF) enables a dynamic fusion process, allowing users to specify their needs through textual descriptions and tailor fused images, offering advantages in tasks that require detailed object-level analysis. A few CIF models have been proposed. For example, Text-Fusion applies adaptive fusion to highlight user-specified semantic objects [4], demonstrating that even with known object locations, varying fusion levels improve downstream task accuracy. Similarly, TeRF enables region-level highlighting based on text [42]. However, these methods can only highlight semantic-level objects, not specific instances. Instance-level CIF remains a challenge.

Our Model. We propose a weakly-supervised instance-level CIF model that highlights user-specified instances.

2.2. Referring Image Segmentation

Referring Image Segmentation (RIS) focuses on segmenting specific objects within an image based on text descriptions [24, 29, 34]. While RIS has advanced significantly for natural images [13], leveraging RGB cues like color, texture, and shape, it faces challenges with IR or VIS images.

(1) Modality Differences. IR images capture thermal radiation but lack RGB's texture and color [20], making object distinction difficult. VIS images provide details but struggle in low-light conditions, where key targets may be obscured [10, 43]. (2) Reduced Instance Differentiation. RIS models excel with natural images by capturing semantic and instance-level details. However, in VIS-IR images, instance-level features are reduced, making it hard to distinguish multiple instances of the same class. For instance, a model may recognize "vehicles" but struggle to distinguish individual instances [4, 42], due to limited visual details.

To address these issues, we design a weakly-supervised instance localization method tailed for VIS-IR images.

3. Method

3.1. Overview

Fig. 2 shows our model with two stages: pseudo-label generation in a weakly-supervised manner and fusion network training. Stage I is the core of our approach, establishing a pixel-wise association between the image content and the input referring text without requiring pixel-level annotations. This stage includes the TIRN and the ISM. Once trained, the model, with frozen parameters, generates reliable masks serving as pseudo-labels for stage II. In stage II, the fusion network takes VIS-IR images and text as inputs, using the generated pseudo-labels as supervision and location references. By defining different learning objectives for target and non-target regions, the network learns to highlight specified instances based on user input.

Stages I and II are independently trained, each with a separate network and no shared modules or parameters. Stage I generates pseudo-labels for stage II, serving as training supervision. During inference (fusion), only the image fusion network is executed, as it has already learned both instance localization and fusion based on textual input.

3.2. Text-to-Image Response Network (TIRN)

As the first part of pseudo-label generation, TIRN'goal is to assign high response values to all objects in multi-source images that match the semantics of the input text, while suppressing responses to irrelevant instances. TIRN enables the localization of semantic-level objects, providing a foundation for localizing instance-level objects in next step.

Inference. TIRN takes a pair of infrared image I_A, visible light image I_B, and user-provided text T as inputs. The inference workflow is illustrated in Fig. 2. First, I_A, I_B, and T are fed into image and text feature extractors, producing vision feature maps F_V^A, $F_V^B \in \mathbb{R}^{H\times W\times C_d}$ and text features $F_T \in \mathbb{R}^{1\times C_d}$. Next, these features are processed through the proposed MFA (detailed in Section 3.4), generating aligned features F_V' and F_T'. Note that F_V' integrates both F_V^A and F_V^B. Then, each pixel of F_V' is multiplied with each query from F_T' to produce a response map $R \in \mathbb{R}^{H\times W}$. The response between the i-th flattened pixel of F_V' and the j-th query of F_T' is computed as follows:

$$R_{i,j} = \sum_{z=1}^{C_d} {F'}_V(i,z)\cdot {F'}_T(j,z). \tag{1}$$

A higher response value in R indicates that the pixel is more likely to belong to the target semantic object, and vice versa. Finally, the average response S_x for each instance x in the instance segmentation map (denoted by S) of OMG-Seg [21] is calculated using Eq. (2):

$$S_x = \frac{\sum_{p\in P_x} R(p)}{|P_x|}, \tag{2}$$

where P_x represents the set of pixels forming instance x, and $|P_x|$ is the total number of pixels in instance x. Noting that most irrelevant instances yield small non-zero response values, we set a threshold 0.1. Any S_x values below this threshold are set to 0. The remaining instances with non-zero S_x values form the set of semantic-level objects P:

$$P = \{x|S_x \neq 0\}. \tag{3}$$

Thus, TIRN achieves localization of semantic-level objects.

Training. Inspired by [25], we adopt a classification-based approach, allowing the model to learn how to select the positive text from a set of positive and negative texts, optimizing the text-to-image response map to localize semantic-level objects. Specifically, we use text and text-image manifold similarity as joint supervision signals to perform this classification. For each pair of multi-source images I_A and I_B, we construct positive text and negative text samples. Positive texts describe the objects in I_A and I_B, while negative samples describe objects in other images. The input includes I_A, I_B, and a positive text Q^p, along with N randomly sampled negative texts Q^N. For each text query Q_j, we compute its image-level score g_j:

$$g_j = \max_i R^a_{i,j} + \frac{1}{HW}\sum_{i=1}^{HW} R^a_{i,j} + \psi(R^a_{:,j}), \tag{4}$$

where $\psi(R^a_{:,j})$ is a regularization term [35]. A higher score indicates a better match between the text and the images. The model suppresses weights for negative texts and enhances those for positive texts by the loss function:

$$L_{TIRN} = L_{cls} + \lambda L_{ms}. \tag{5}$$

L_{cls} is the standard classification loss, described by:

$$L_{\text{cls}}(g,z) = \frac{1}{N+1}\sum_{j=1}^{1+N} z_j \log\left(\frac{1}{1-e^{-g_j}}\right) + (1-z_j)\log\left(\frac{e^{-g_j}}{1+e^{-g_j}}\right), \tag{6}$$

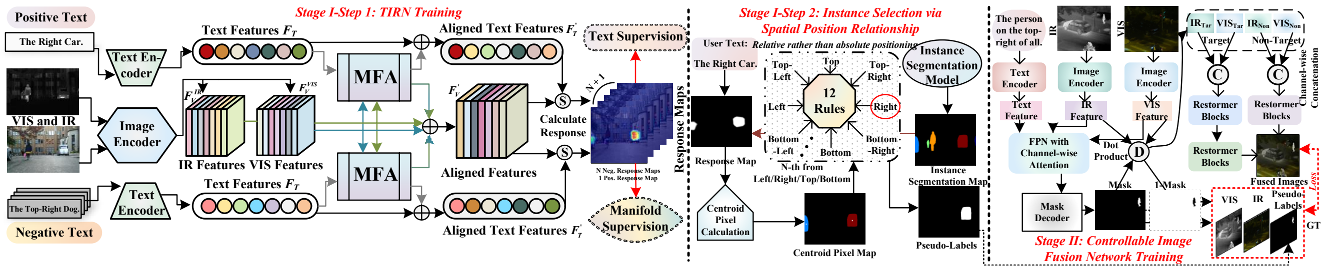

Figure 2. Schematic diagram of the proposed model. During training, our model operates in two stages: pseudo-label generation and fusion network training. The first stage consists of two main steps: text-to-image response (TIRN) and instance selection.

where $z \in \mathbb{R}^{1\times(1+N)}$ represents supervision signals, with 1 indicating positive text and 0 indicating negative text. L_{ms} first computes the manifold similarity M between matched texts and images in a well-trained feature space. It then uses this similarity as a supervision signal to adjust the attention map A in MFA, supported by manifold priors, as follows:

$$L_{ms}=\alpha\sum_{i,j}\|A_{i,j}-f(M_{i,j})\|_2-\beta\sum_{i,j}\|\mathrm{Var}(A_{i:})-\mathrm{Var}(M_{i:})\|_2, \quad (7)$$

where f represents sigmoid; details on computing manifold similarity and distances are provided in Section 3.4. The first term ensures consistency between A and $f(M)$, guiding attention distribution according to manifold structure, while the second term prevents over-reliance on manifold similarity, ensuring diverse feature representation. As this loss decreases, text tokens with high manifold similarity to image patches receive greater attention, improving sensitivity to positive samples while distinguishing negative ones.

3.3. Instance Selection Module (ISM)

ISM selects the user-specified instance from the set P. Given that distinguishing between instance-level objects in infrared or visible images is challenging, we use the relative positional relationships provided in text as the basis for selection. First, the module computes each instance's centroid pixel in P by averaging its pixel coordinates, as follows:

$$h_i = (\frac{1}{n}\sum_{j=1}^{n} x_j, \frac{1}{n}\sum_{j=1}^{n} y_j), \quad (8)$$

where n represents the total number of pixels in the i-th instance, and (x_j, y_j) denotes the pixel coordinates. This results in a set of centroid coordinates $H = \{h_1, h_2, \dots, h_k\}$, where k is the number of instances. Next, the model identifies position-related keywords, such as "left" or "bottom-right." Finally, the module compares the centroid coordinates of each instance and selects the target instance h^* according to 12 designed rules, as described by Eq. (9).

$$h^* = \mathcal{R}_i(D_i), \quad (9)$$

where D_i denotes the detected spatial term in the vocabulary set D, and R_i represents the designed spatial relationship rule for D_i. D includes 12 position-related terms: left, right, top, bottom, top-left, top-right, bottom-left, bottom-right, and the N-th from the left/right/top/bottom. The first eight words apply when the referenced object's position is easily described, while the last four words address more complex cases, such as selecting an intermediate object from multiple instances. Each term corresponds to a strict, custom instance selection rule (details in appendix). By using relative positioning instead of absolute positioning, any instance can be accurately referenced.

This process shows that ISM selects the user-specified instance from multiple candidates without extra training, generating pseudo-labels for fusion network training.

3.4. Manifold-Based Feature Alignment Module

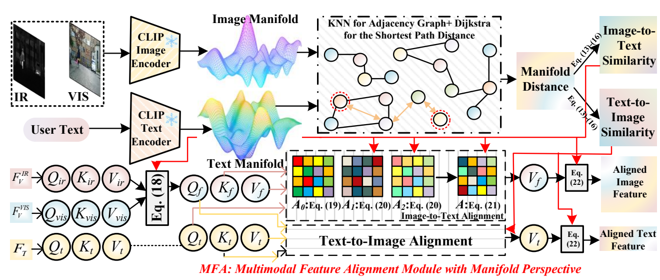

Figure 3. Schematic diagram of MFA. Manifold prior supports it.

To align multimodal feature in TIRN, we design MFA from a manifold perspective [14, 36, 47, 48]. MFA treats the three inputs—VIS image, IR image, and text—as separate manifolds. Supported by discovered manifold prior (see Fig. 1), the core design leverages manifold similarity as a guiding prior for alignment. MFA comprises three stages: computing manifold similarity, joint VIS-IR representation, and bidirectional alignment between text and image.

The first stage calculates the similarity matrices required for later steps. The second stage combines complementary features from VIS and IR images, avoiding mismatches between image objects and text caused by the limitations of a single source image, as discussed in Section 2.2. The final stages focus on aligning text and image features bidirectionally. The inputs include patch embeddings Q_{vis}, K_{vis}, V_{vis}

from F_V^{vis}; Q_{ir}, K_{ir}, V_{ir} from F_V^{ir}; and Q_t, K_t, V_t from text tokens, outputting aligned features F_V' and F_T'.

Manifold Similarity Computation. We compute manifold similarity matrices: M_{vis-ir} for VIS-IR alignment, and $M_{t\to v}$, $M_{v\to t}$ for text-image alignment. Due to space constraints, we take the computation of $M_{t\to v}$ as an example. We first extract features from I_A, I_B, and T using feature extractors, ensuring the effectiveness of manifold structure. Let the image features be $\mathbf{V} \in \mathbb{R}^{N\times d}$ and text features $\mathbf{T} \in \mathbb{R}^{M\times d}$, where N and M represent the number of patches and tokens, respectively. These features are mapped into a shared space and concatenated into $\mathbf{X}$:

$$\mathbf{X} = [\mathbf{V}; \mathbf{T}] \in \mathbb{R}^{(N+M)\times d}. \tag{10}$$

Next, we construct a KNN graph G from $\mathbf{X}$, capturing local relationships between features:

$$G = KNN(\mathbf{X}, k). \tag{11}$$

The manifold distance matrix $d_{t\to v} \in \mathbb{R}^{M\times N}$ is then computed, where $d_{t\to v}(j,i)$ represents the shortest path between text feature $\mathbf{T}_j$ and image feature $\mathbf{V}_i$ calculated by the Dijkstra algorithm [7] denoted by Γ:

$$d_{t\to v}(j,i) = \Gamma(G, N+j, i). \tag{12}$$

Then, we convert distances into similarity using RBF [2]:

$$M_{t\to v}^{fixed}(j,i) = \exp(-\frac{d_{t\to v}(j,i)^2}{2\sigma^2}), \tag{13}$$

where σ is the RBF scale parameter, controlling similarity. In order to prevent the fixed similarity matrix from hindering the training of the network, we perform an affine transformation on $M_{t\to v}^{fixed}$ to better fit the network:

$$M_{t\to v}^{learned} = W_0 \times M_{t\to v}^{fixed} + b, \tag{14}$$

where W_0 denotes the learnable matrix; b represents the learnable parameter. Finally, $M_{t\to v}^{*}$ is a linear combination of $M_{t\to v}^{fixed}$ and $M_{t\to v}^{learned}$ based on learnable parameter λ:

$$M_{t\to v}^{*} = \lambda \cdot M_{t\to v}^{fixed} + (1-\lambda)\cdot M_{t\to v}^{learned}. \tag{15}$$

Finally, to align with subsequent computations, $M_{t\to v}^{*} \in \mathbb{R}^{M\times N}$ is reshaped through interaction with $\mathbf{T} \in \mathbb{R}^{M\times d}$, producing $M_{t\to v} \in \mathbb{R}^{N\times d}$, as follows:

$$M_{t\to v} = (M_{t\to v}^{*})^T \times \mathbf{T}. \tag{16}$$

$M_{t\to v}$ guides the alignment from text to image. The computations of $M_{v\to t}$ and M_{vis-ir} follow a similar process.

Joint VIS-IR Representation. To obtain a comprehensive vision representation, we first globally pool M_{vis-ir} to derive M_{gap}. Then, linear projections generate parameters:

$$\alpha = \sigma(W_\alpha \times M_{\text{gap}} + b_\alpha), \quad \beta = \sigma(W_\beta \times M_{\text{gap}} + b_\beta), \tag{17}$$

where σ denotes the sigmoid function. Finally, the fused Q_f, K_f, and V_f are computed as:

$$Q_f = \alpha Q_{ir} + \beta Q_{vis}, K_f = \alpha K_{ir} + \beta K_{vis}, V_f = \alpha V_{ir} + \beta V_{vis}. \tag{18}$$

Text-to-Image Alignment. The process first computes the attention scores between $Q_t \in R^{M\times d}$ and $K_f \in R^{N\times d}$:

$$A_0 = \frac{Q_t \times K_f^T}{\sqrt{d_k}}. \tag{19}$$

Then, $M_{t\to v}$ interacts with Q_t and K_f, respectively:

$$A_1 = \frac{Q_t \times M_{t\to v}^T}{\sqrt{d_m}}, \quad A_2 = \frac{M_{t\to v} \times K_f^T}{\sqrt{d_m}}. \tag{20}$$

This step uses manifold similarity as a prior to guide attention toward potential text-image correspondences, enhancing feature alignment. These three attention maps are fused:

$$A = A_0 + \tau_1 \cdot A_1 + \tau_2 \cdot (W_1 \times A_2), \tag{21}$$

where τ_1 and τ_2 are learnable parameters that balance the importance of the three terms; $W_1 \in R^{M\times N}$ denotes the learnable matrix. After applying Softmax to normalize A, the weights are used to compute the output:

$$F_T^{'} = A \times V_f + \gamma \cdot (M_{t\to v} \times V_f). \tag{22}$$

The learnable γ ensures that $M_{t\to v}$ continues to impact output. This output serves as the aligned text features $F_T^{'}$.

Image-to-Text Alignment. This process mirrors text-to-image alignment, with the following differences: (1) Q is derived from Q_f, while K and V are K_t and V_t; (2) it uses the manifold similarity $M_{v\to t}$. The output serves as the aligned image features $F_V^{'}$.

3.5. Image Fusion Network

In stage II, localization maps from stage I serve as pseudo-labels to train the fusion network. Training and inference are given here, with architecture presented in appendix.

Training. The network takes VIS, IR, and user-provided text as inputs, using pseudo-labels (binary masks) as location references. It employs two fusion modes to highlight target regions. For ROI in the pseudo-labels, effective features from both source images serve as supervision signals, including the brightness from IR and color and texture from VIS. To achieve this, we define the loss function with L_1:

$$\begin{aligned} L_{\text{ROI}} = &\varpi_1 L_1(Y_{FI}; IR) + \varpi_2 L_1(CB_{FI}; CB_{VIS}) + \\ &\varpi_3 L_1(CR_{FI}; CR_{VIS}) + \varpi_4 L_1(G_{FI}; G_{VIS}), \end{aligned} \tag{23}$$

where Y_{FI}, CB_{FI}, and CR_{FI} are luminance and chrominance channels of fused images (via YCBCR decomposition); G denotes gradient calculation. 1st term enhances target instance brightness with IR, 2nd and 3rd terms ensure accurate color, and 4th term preserves textures, ensuring that highlights the object without compromising detail.

For non-target regions, we follow standard practices, using both source images as supervision signals:

$$L_{\text{non-ROI}} = \omega_1 L_1(FI; IR, VIS) + \omega_2 L_{ssim}(FI; IR, VIS). \tag{24}$$

In addition to L_1, we use SSIM [45] in non-target regions.

We compute separate losses for ROI and non-ROI regions with pseudo-labels. To minimize loss, the network learns both fusion modes while localizing based on text.

Inference. The fusion network ζ_θ takes VIS images, IR images, and text as inputs and generates fused images FI:

$$FI = \zeta_\theta(IR, VIS, T). \tag{25}$$

4. Experimental Results

4.1. Setting

In stage 1, we set $\lambda = 500$, $\alpha = 1$, $\beta = 100$ and resize input images to 320 × 320. We use AdamW (initial learning rate 5e^{-5}) for 120 epochs, batch size 12. In stage 2, $\omega_1 = 50$, $\omega_2 = 4$, $\varpi_1 = 1$, $\varpi_2 = 2$, $\varpi_3 = 2$, $\varpi_4 = 5$. We use Adam with a learning rate of 1e^{-4}, training for 100 epochs with batch size 2. Experiments are conducted on a machine with an RTX 4090 GPU, i9-14900K CPU, 128GB RAM, and PyTorch 2.3.0.

4.2. Dataset and Metrics

Both TIRN and the fusion network require training datasets. For TIRN, we use Li et al.'s dataset [4] with 2000 VIS-IR image pairs, each paired with three text descriptions referring to semantic-level objects. The fusion network's training set includes source image pairs, instance-level text, and pseudo-labels generated in our model's first stage, with restructured text descriptions.

We select LLVIP [12], M^{3}FD [26], MSRS [38], and TNO [41] as test sets, which are widely used in IF field. For evaluating controllable image fusion (CIF) models, each image pair is given a text description that randomly refers to an object in the image. All CIF models use the same image-text pairs for fair fusion results.

We use the metrics Q_{CB} [3], Q_Y [18], Q_E [33], Q_W [33], SF [45], AG [6], VIF [9] and $Q_{AB/F}$ [49]. SF and AG are unsupervised metrics; the rest are supervised. For supervised metrics, we follow convention by using each source image as ground truth for metric calculation.

4.3. Comparison with SOTA

Comparison Models. We select ten methods, including FusionGAN [31], DIDFuse [55], GANMcC [32], DeFusion [23], ReCoNet [11], CDDFuse [56], DDFM [57], LRRNet [16], FILM [59], and TextFusion [4]. The first nine models are uncontrollable fusion (UIF) methods, using only source images as input; while TextFusion and our model are controllable (CIF), using both source images and text inputs.

Qualitative Evaluation. Fig. 4 illustrates the qualitative comparison across three samples from different test sets. Both our model and TextFusion adjust the fused image based on text. Our model uniquely highlights instance-level objects, whereas TextFusion emphasizes all objects matching the text. For example, in the first group, TextFusion fails to accurately isolate the instances described by "the car on the right" and "the person on the top left," while our model precisely localizes the targets. "No text" means no text input, where our model applies only the non-target fusion mode, achieving promising results. This demonstrates its strong performance in static fusion.

Quantitative Evaluation. Table 1 reports the results of various models across eight metrics. Our model achieves superior results in most metrics, showing its effectiveness in instance localization and fusion quality.

4.4. Ablation Study

As our main innovation focuses on stage I, we measure response maps or pseudo-labels using segmentation metrics, with segmentation maps of referred instances from large models serving as GT. Quantitative results are presented, with qualitative results in appendix.

Effect of TIRN. TIRN is trained with text and manifold similarity supervision. To assess its impact, we conduct the following ablations: (1) removing text supervision; (2) removing the manifold similarity; (3) removing TIRN entirely; (4) the default setup. Table 2(I) shows that both supervision signals significantly improve model performance. TIRN is crucial for generating accurate pseudo-labels, as its removal results in a marked drop in pseudo-label accuracy.

Effect of MFA. To evaluate MFA, we perform the following ablations: (1) removing the module; (2) replacing it with cross-attention; (3) using random matrices instead of the manifold similarity matrix; (4) using well-trained CLIP instead of MFA; (5) the default setup. Results in Table 2(II) confirm that MFA, particularly the manifold similarity, is vital for enhancing model performance.

Effect of Instance Selection via Spatial Relationships. This module refines instance localization using positional descriptions. We perform: (1) removing the module; (2) reducing spatial rules from 12 to 4; (3) the default setup. Table 2(III) shows that without this module, pseudo-labels fail to capture instance-level objects, and fewer rules reduce localization accuracy for positions like "top-left".

4.5. Validation of the Necessity of Tailored Instance Localization Method for VIS-IR

As discussed in Sections 1 and 2.2, RIS models designed for natural images are unsuitable for object localization in VIS-IR. To verify, we test several RIS models (LAVT [50], VLT [8], MG [5]) on IR and VIS images. Unlike these models, which process only single images and yield suboptimal results due to modality limitations, our VIS-IR-tailored method takes two images as input to create a joint visual representation, reducing localization difficulty. As shown

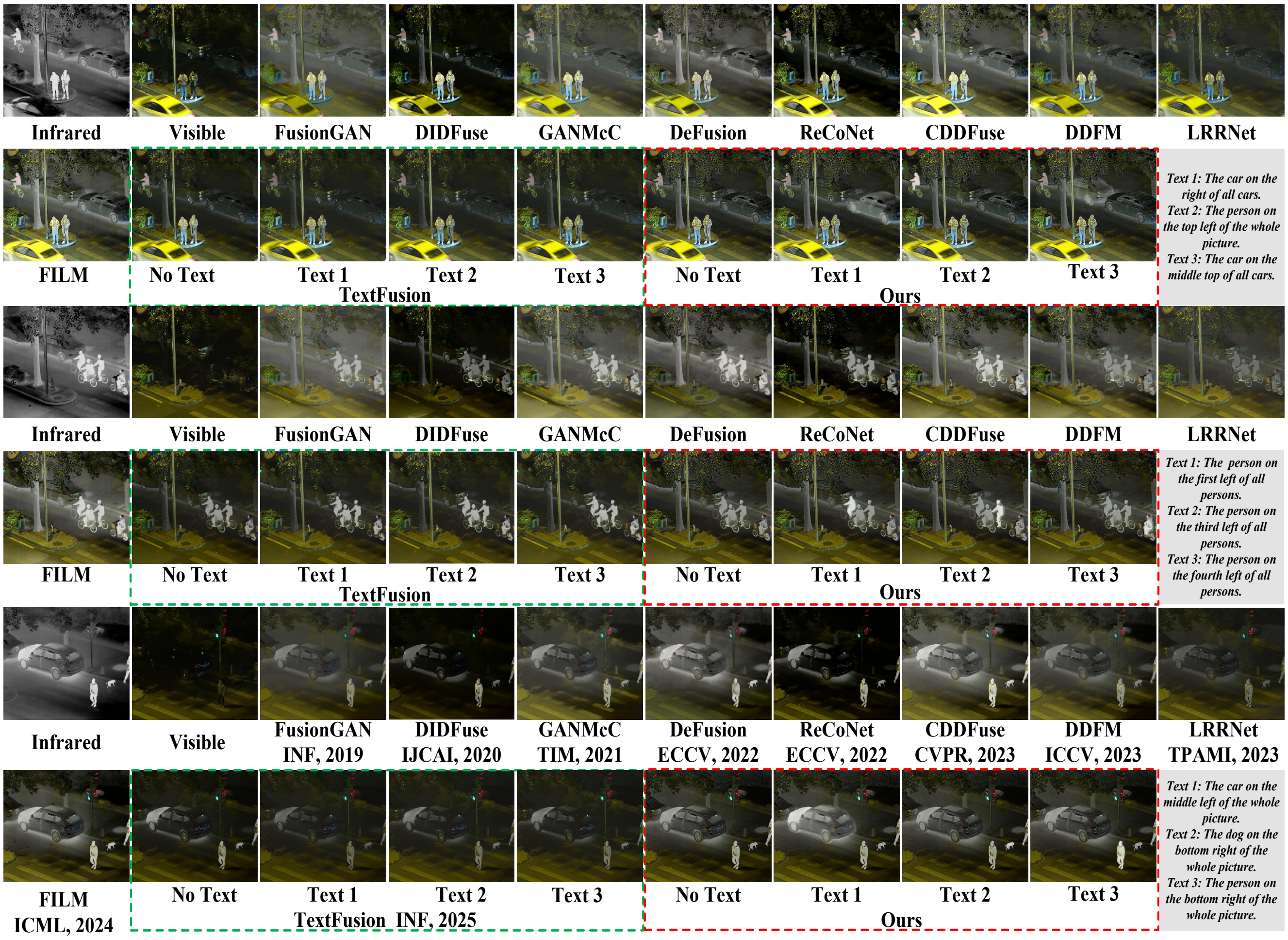

Figure 4. Qualitative comparison of fusion models. The first nine UIF models yield static results regardless of text. TextFusion and ours take one text and two source images as input. TextFusion highlights semantic-level objects, while our model targets the referenced instance.

Table 1. Quantitative comparison of various fusion models across four test sets.

Test Sets	Methods	$Q_{CB}\uparrow$	$Q_Y\uparrow$	$Q_E\uparrow$	$Q_W\uparrow$	$SF\uparrow$	$AG\uparrow$	$VIF\uparrow$	$Q_{AB/F}\uparrow$	Test Sets	Methods	$Q_{CB}\uparrow$	$Q_Y\uparrow$	$Q_E\uparrow$	$Q_W\uparrow$	$SF\uparrow$	$AG\uparrow$	$VIF\uparrow$	$Q_{AB/F}\uparrow$
	FusionGAN	0.251	0.487	0.306	0.289	5.970	1.609	0.498	0.212		FusionGAN	0.330	0.552	0.376	0.333	8.012	2.688	0.388	0.264
	DIDFuse	0.424	0.338	0.432	0.389	10.313	2.098	0.444	0.267		DIDFuse	0.445	0.653	0.668	0.654	15.423	5.306	0.669	0.511
	GANMcC	0.326	0.572	0.368	0.413	6.286	1.874	0.654	0.292		GANMcC	0.392	0.615	0.409	0.431	7.590	2.695	0.527	0.289
	DeFusion	0.432	0.734	0.576	0.650	8.940	2.523	0.776	0.466		DeFusion	0.427	0.672	0.502	0.467	8.408	2.945	0.549	0.345
	ReCoNet	0.407	0.395	0.611	0.551	9.915	2.774	0.580	0.408		ReCoNet	0.455	0.740	0.706	0.670	11.955	4.479	0.604	0.508
LLVIP	CDDFuse	0.459	0.857	0.767	0.762	13.278	3.432	0.958	0.641	M^3FD	CDDFuse	0.475	0.870	0.793	0.722	16.491	5.417	0.781	0.632
	DDFM	0.406	0.613	0.370	0.451	5.785	1.869	0.714	0.300		DDFM	0.405	0.656	0.555	0.532	9.725	3.379	0.606	0.449
	LRRNet	0.360	0.521	0.512	0.411	8.874	2.277	0.558	0.405		LRRNet	0.435	0.721	0.646	0.574	26.690	9.269	0.762	0.461
	FILM	0.485	0.759	0.825	0.791	14.361	3.926	0.976	0.675		FILM	0.493	0.839	0.816	0.748	16.757	5.545	0.806	0.653
	TextFusion	0.455	0.692	0.654	0.552	11.765	2.894	0.738	0.495		TextFusion	0.477	0.713	0.755	0.662	16.836	5.523	0.619	0.544
	Ours	0.511	0.784	0.843	0.803	14.572	4.005	0.996	0.711		Ours	0.517	0.905	0.833	0.754	16.627	5.609	0.838	0.694
	FusionGAN	0.322	0.391	0.206	0.206	4.354	1.446	0.442	0.140		FusionGAN	0.408	0.539	0.386	0.373	6.269	2.362	0.418	0.224
	DIDFuse	0.407	0.238	0.432	0.402	9.644	2.006	0.304	0.202		DIDFuse	0.463	0.655	0.602	0.614	11.768	4.249	0.593	0.403
	GANMcC	0.424	0.574	0.404	0.444	5.664	1.999	0.635	0.302		GANMcC	0.437	0.607	0.420	0.445	6.217	2.513	0.513	0.275
	DeFusion	0.514	0.749	0.730	0.753	8.146	2.644	0.730	0.507		DeFusion	0.482	0.707	0.592	0.568	6.598	2.675	0.553	0.359
	ReCoNet	0.378	0.347	0.720	0.692	9.975	2.990	0.490	0.404		ReCoNet	0.463	0.673	0.592	0.610	7.958	3.353	0.531	0.373
MSRS	CDDFuse	0.567	0.827	0.868	0.859	11.556	3.734	1.051	0.693	TNO	CDDFuse	0.450	0.787	0.697	0.659	11.621	4.330	0.730	0.496
	DDFM	0.482	0.662	0.582	0.579	7.388	2.513	0.743	0.474		DDFM	0.437	0.485	0.473	0.484	8.128	3.213	0.371	0.292
	LRRNet	0.393	0.507	0.659	0.632	8.473	2.641	0.541	0.454		LRRNet	0.483	0.701	0.513	0.513	9.438	3.627	0.538	0.367
	FILM	0.569	0.822	0.873	0.863	11.726	3.858	1.056	0.723		FILM	0.483	0.821	0.745	0.726	12.579	4.556	0.725	0.529
	TextFusion	0.490	0.618	0.795	0.805	10.230	3.045	0.694	0.475		TextFusion	0.525	0.750	0.581	0.582	10.217	3.986	0.598	0.392
	Ours	0.570	0.875	0.887	0.876	11.788	3.877	0.997	0.698		Ours	0.520	0.829	0.746	0.733	12.935	4.791	0.704	0.566

in Fig. 5, these models fail to localize instances correctly, while our method (Stage I) produces accurate results. Numerical results are provided in supplementary material.

To show the impact of MFA, the right part of Fig. 5 visualizes TIRN outputs using MFA versus cross-attention, with MFA yielding higher response values for correct ob-

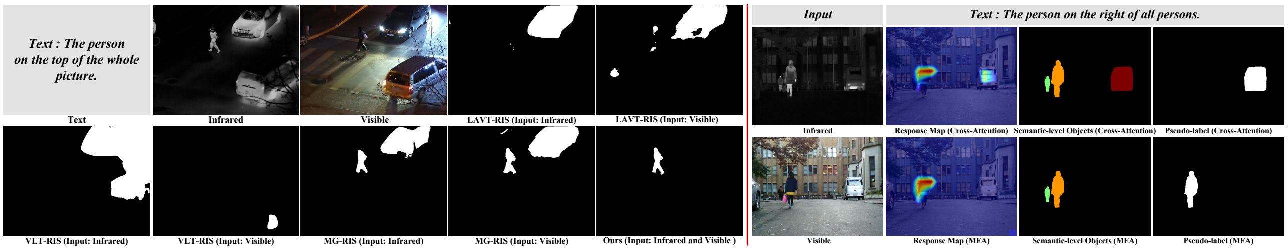

Figure 5. Visualization. The left part shows a comparison of our localization method with recent RIS models on VIS and IR images. The right part illustrates the impact of our MFA and cross-attention on pseudo-label accuracy.

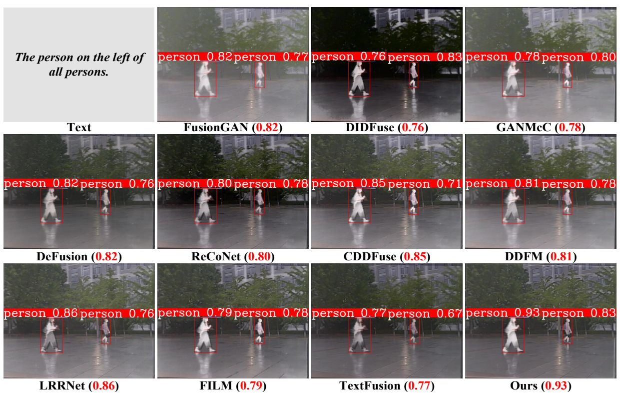

Figure 6. Qualitative comparisons of different fusion models on the TOD task. Fused images generated by our model enable YOLO-v5 to achieve higher confidence scores for target instances.

Table 2. Quantitative results of multiple ablation experiments. The checkmark indicates which of RM or PL is used to calculate metrics. MS: Manifold Similarity; CA: Cross Attention; RM: Response Map; PL: Pseudo-Label; SR: Spatial Relationships. PointIt and PointM can only be used for the evaluation of RM.

	Descriptions	RM	PL	Metrics			
				P@0.5[37]	mIoU	PointIt[25]	PointM[25]
I	w/o Text-Supervision	✓		0.240	1.231	9.886	6.027
	w/o MS	✓		5.941	22.257	59.073	44.274
	Default Setup	✓		23.437	41.714	83.394	75.526
II	w/o MFA	✓		4.985	21.052	57.959	42.252
	MFA→CA	✓		9.526	25.615	65.606	50.311
	MS→Random Matrix	✓		2.258	11.790	21.974	13.141
	MFA→CLIP	✓		4.661	20.713	56.816	41.227
	Default Setup	✓		23.437	41.714	83.394	75.526
III	w/o SR	✓		23.437	41.714	83.394	75.526
	w/o TIRN		✓	32.283	30.212	-	-
	Rules Reduce: 8→4		✓	60.629	54.579	-	-
	Default Setup		✓	73.228	66.065	-	-

jects. It also shows the two-step pseudo-label generation: TIRN captures all semantic-level objects, and then instance selection module pinpoints the correct instance.

4.6. Downstream Task: Targeted Object Detection

Our model enhances TOD by highlighting referenced instances. To assess the impact of different fusion models on TOD, we use YOLO-v5 to detect objects in various fused images (see Fig. 6, Table 3). By emphasizing low-

Table 3. Comparison results of fusion models on the TOD task.

Fusion Methods	Recall	mAP		
		@0.50	@0.75	@[0.5:0.95]
FusionGAN [31]	0.207	0.307	0.179	0.170
DIDFuse [55]	0.272	0.402	0.250	0.231
GANMcC [32]	0.267	0.393	0.235	0.222
DeFusion [23]	0.259	0.378	0.233	0.214
ReCoNet [11]	0.287	0.423	0.252	0.243
CDDFuse [56]	0.292	0.430	0.261	0.246
DDFM [57]	0.281	0.422	0.245	0.234
LRRNet [16]	0.304	0.447	0.260	0.252
FILM [59]	0.300	0.45	0.261	0.251
TextFusion [4]	0.288	0.423	0.253	0.243
Ours	0.314	0.472	0.268	0.263

brightness instances, our model improves YOLO-v5's confidence scores and achieves superior quantitative results on the M^3FD dataset. Compared to other fusion models, its instance-level controllability offers distinct advantages in TOD, further underscoring its value.

5. Conclusion

We propose a novel instance-level controllable image fusion model for VIS-IR fusion, enabling users to highlight any instance via text. Lacking suitable datasets and labels, we employ a two-stage weakly supervised approach to generate pseudo-labels. In stage I, we train TIRN with text and manifold similarity as joint supervision, using MFA to align image patches with text tokens and applying twelve spatial rules for instance selection. In stage II, the fusion network locates target objects based on text input and applies two fusion modes. Results show accurate instance highlighting and state-of-the-art fusion performance by our model.

Acknowledgement

This work was supported by the National Natural Science Foundation of China [No.62401097]; and the Natural Science Foundation of Liaoning Province (Doctoral Research Start-up Project) [No.2024-BS-028]; and the Fundamental Research Funds for the Central Universities (No.0854-53).

References

[1] Gaurav Bhatnagar and Zheng Liu. A novel image fusion framework for night-vision navigation and surveillance. *Signal, Image and Video Processing*, 9:165–175, 2015. 1

[2] Martin Dietrich Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2000. 5

[3] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image and vision computing*, 27(10):1421–1432, 2009. 6

[4] Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, Zhangyong Tang, and Josef Kittler. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117:102790, 2025. 1, 2, 3, 6, 8

[5] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024. 6

[6] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. 6

[7] Edsger W Dijkstra. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*, pages 287–290. 2022. 5

[8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916, 2022. 6

[9] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information fusion*, 14(2):127–135, 2013. 6

[10] Xinyu Hu, Yang Liu, and Feng Yang. Pfcfuse: A poolformer and cnn fusion network for infrared-visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 2024. 3

[11] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, pages 539–555. Springer, 2022. 6, 8

[12] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 6

[13] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021. 3

[14] Ron Kimmel, Nir Sochen, and Ravi Malladi. From high energy physics to low level vision. In *Scale-Space Theory in Computer Vision: First International Conference, Proceedings 1*, pages 236–247. Springer, 1997. 2, 4

[15] Guofa Li, Yongjie Lin, and Xingda Qu. An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Information Fusion*, 71:109–129, 2021. 2

[16] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023. 6, 8

[17] Huafeng Li, Junzhi Zhao, Jinxing Li, Zhengtao Yu, and Guangming Lu. Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. *Information Fusion*, 95:26–41, 2023. 2

[18] Shanshan Li, Richang Hong, and Xiuqing Wu. A novel similarity based quality metric for image fusion. In *2008 International Conference on Audio, Language and Image Processing*, pages 167–172. IEEE, 2008. 6

[19] Xingyuan Li, Yang Zou, Jinyuan Liu, Zhiying Jiang, Long Ma, Xin Fan, and Risheng Liu. From text to pixels: a context-aware semantic synergy solution for infrared and visible image fusion. *arXiv preprint arXiv:2401.00421*, 2023. 2

[20] Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pages 270–288. Springer, 2024. 3

[21] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27948–27959, 2024. 3

[22] Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7534–7544, 2025. 2

[23] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022. 6, 8

[24] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017. 3

[25] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023. 2, 3, 8

[26] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 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*, pages 5802–5811, 2022. 1, 6

[27] Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized

semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 2024. 2

[28] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5):1748–1775, 2024. 1

[29] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021. 3

[30] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 2

[31] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019. 6, 8

[32] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020. 6, 8

[33] Gemma Piella and Henk Heijmans. A new quality metric for image fusion. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, pages III–173. IEEE, 2003. 6

[34] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 3

[35] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 3

[36] Nir Sochen, Ron Kimmel, and Ravi Malladi. A general framework for low level vision. *IEEE transactions on image processing*, 7(3):310–318, 1998. 2, 4

[37] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23570–23580, 2023. 8

[38] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 6

[39] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. 1

[40] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25:5413–5428, 2022. 1

[41] Alexander Toet and Maarten A Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):010901–010901, 2012. 6

[42] Hebaixu Wang, Hao Zhang, Xunpeng Yi, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. Terf: Text-driven and region-aware flexible visible and infrared image fusion. In *ACM Multimedia 2024*. 2, 3

[43] Jing Wang, Long Yu, and Shengwei Tian. Cross-attention interaction learning network for multi-model image fusion via transformer. *Engineering Applications of Artificial Intelligence*, 139:109583, 2025. 3

[44] Xuan Wang, Kaiqiang Li, and Abdellah Chehri. Multi-sensor fusion technology for 3d object detection in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[46] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2

[47] Zeyu Wang, Xiongfei Li, Libo Zhao, Haoran Duan, Shidong Wang, Hao Liu, and Xiaoli Zhang. When multi-focus image fusion networks meet traditional edge-preservation technology. *International Journal of Computer Vision*, 131(10): 2529–2552, 2023. 4

[48] Zeyu Wang, Libo Zhao, Jizheng Zhang, Rui Song, Haiyu Song, Jiana Meng, and Shidong Wang. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, pages 4646—-4668, 2025. 4

[49] Costas S Xydeas, Vladimir Petrovic, et al. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000. 6

[50] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 6

[51] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 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*, pages 27026–27035, 2024. 2

[52] Xunpeng Yi, Yong Ma, Yansheng Li, Han Xu, and Jiayi Ma. Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation*, 6(4), 2025. 1

[53] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[54] Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[55] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image de-

composition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020. 6, 8

[56] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. 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*, pages 5906–5916, 2023. 6, 8

[57] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023. 6, 8

[58] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024. 1

[59] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, PENG WANG, et al. Image fusion via vision-language model. In *Forty-first International Conference on Machine Learning*, 2024. 6, 8

[60] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. *arXiv preprint arXiv:2505.19858*, 2025. 2

[61] Naishan Zheng, Man Zhou, Jie Huang, and Feng Zhao. Frequency integration and spatial compensation network for infrared and visible image fusion. *Information Fusion*, 109: 102359, 2024. 2

[62] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. 2

[63] Yang Zou, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Enhancing neural radiance fields with adaptive multi-exposure fusion: A bilevel optimization approach for novel view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7882–7890, 2024. 2