ControlFuse: Instruction-guided Multi-Granularity Controllable Image Fusion

Libo Zhao[1], **Xiaoli Zhang**[1], **Zeyu Wang**[2*]

[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]College of Computer Science and Engineering, Dalian Minzu University, Dalian, China
zhaolb22@mails.jlu.edu.cn, wangzeyu@dlnu.edu.cn

Abstract

Infrared and Visible Image Fusion (IVIF) produces enhanced images by fusing complementary visual information. However, most existing methods generate fixed outputs and cannot flexibly adapt to user-specific requirements. Recent text-guided approaches offer partial control but are limited to global or semantic levels, lacking instance-level control. This limitation arises from two challenges: first, the lack of datasets that directly link textual instructions with corresponding spatial annotations, and second, the use of coarse cross-modal alignment methods that struggle to precisely match textual instructions with visual features. To overcome these challenges, we propose ControlFuse, a controllable IVIF framework enabling multi-granularity fusion across global, semantic, and instance levels, guided by user instructions. First, we construct an automated multi-granularity dataset that provides explicit textual-mask correspondences at these three levels. Second, inspired by manifold geometry, we design a Multimodal Feature Interaction Module (MFIM) comprising Feature Manifold Converter (FMC) and Curvature-Guided Interaction (CGI). FMC projects textual and visual features into a unified manifold space, while CGI leverages manifold curvature as a geometric cue to refine cross-modal alignment. Extensive experiments validate ControlFuse, outperforming state-of-the-art methods in robustness and flexibility.

Code — https://github.com/zhaolb4080/ControlFuse

Introduction

Infrared imaging records thermal radiation, providing reliable cues in darkness, smoke, and fog, but with low spatial resolution and sparse texture. Visible imaging, acquired through reflected light, offers high resolution and rich colors with detailed edges, yet its quality degrades under poor illumination or adverse weather. The highly complementary natures motivate Infrared and Visible Image Fusion (IVIF) (Liu et al. 2024b; Tang et al. 2025b; Liu et al. 2024a). IVIF integrates infrared and visible data to produce comprehensive fused images, enhancing visual quality and boosting the accuracy of downstream tasks, e.g., semantic segmentation (Liu et al. 2023; Wu et al. 2025), and object detection (Li et al. 2025b; Wang et al. 2025c).

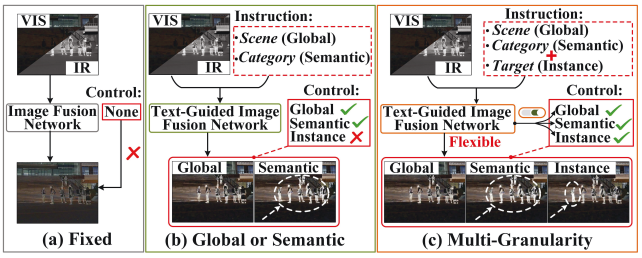

Figure 1: Comparison of controllable IVIF fusion paradigms given the same VIS-IR inputs. (a) Conventional fusion produces a fixed result. (b) Text-guided fusion enables control at the global and semantic levels but does not support instance-level control. (c) Our ControlFuse enables instruction-guided fusion with multi-granularity control at the global, semantic, and instance levels.

To achieve effective fusion, numerous IVIF methods have been proposed over the past decades. Traditional methods, based on information theory, focus on preserving source information but often struggle to optimize fused image quality, especially with redundant features or scenario-specific requirements (Zhang, Ye, and Xiao 2020). Deep learning-based methods (Liu et al. 2025a; Zhao et al. 2023a,c; Huang et al. 2022; Wang et al. 2025a; Liu et al. 2025b), benefiting from powerful representation learning and nonlinear modeling capabilities, markedly surpass traditional approaches in fusion performance. Despite the progress in fusion techniques, most methods produce fixed outputs once trained, lacking controllability to meet varying application-specific requirements. As illustrated in Fig. 1(a), fixed outputs cannot selectively highlight targets, such as adaptively enhancing infrared content to detect pedestrians in low-light conditions, thereby restricting their effectiveness for specific tasks (Tang, Li, and Ma 2025; Tang, He, and Liu 2022; Li et al. 2025a; Yang et al. 2025; Tang, He, and Liu 2024; Li et al. 2024).

To achieve controllable fusion, recent studies have begun integrating textual information into the IVIF framework, enabling dynamic fusion guided by user-specified instructions (Zhao et al. 2024b). However, controllable fusion methods

*Corresponding author.

capable of flexibly highlighting user-specified targets remain unexplored. As illustrated in Fig. 1(b), existing approaches typically provide either global scene enhancement through full scene textual descriptions or semantic-level highlighting guided by category text instructions (Tang et al. 2025a; Yi et al. 2024; Zhang et al. 2025; Cheng et al. 2025). Consequently, these methods lack precise instance-level control, i.e., they cannot individually highlight specific targets.

The limitation primarily stems from two factors. First, existing IVIF datasets usually do not provide detailed annotations. Specifically, they lack clear textual descriptions directly connected to exact locations of individual objects within images. Without these annotations, it is difficult to train fusion models effectively to highlight specific targets at the instance level. Second, current multimodal fusion strategies in IVIF frameworks commonly utilize relatively coarse alignment methods, such as global cross-attention, element-wise product, or straightforward feature concatenation. While these approaches effectively combine multimodal information at a broad level, they generally lack sufficient granularity to precisely align detailed textual instructions with the visual features of specific objects. Consequently, fusion models trained using these coarse methods often struggle to accurately highlight targets specified by users in the resulting fused images. Based on the analysis above, we identify two critical solutions:

For the first limitation, we construct a multi-granularity IVIF dataset using an automated annotation pipeline. This dataset aligns textual instructions with corresponding annotation masks at global, semantic, and instance granularities. For example, global-level annotations correspond to instructions such as "enhance the entire scene," with mask covering all targets present. Semantic-level annotations match instructions like "highlight all pedestrians," accompanied by mask that covers every pedestrian in the scene. Instance-level annotations align precisely with instructions such as "highlight the pedestrian on the left," with mask covering only the specified object. This structured alignment between textual inputs and masks enables flexible fusion directly guided by user instructions.

For the second limitation, we introduce a multimodal feature interaction module designed to precisely align textual instructions with visual features. This module is inspired by the observation that paired text-image data can be viewed as two manifolds, where high manifold curvature corresponds to critical structural elements in each domain. Specifically, in images, high curvature highlights edges and contours essential for object localization; in text, it identifies semantically distinctive terms useful for cross-modal alignment. Accordingly, we propose two complementary components: Feature Manifold Converter (FMC) and Curvature-Guided Interaction (CGI). FMC is a graph-based module optimized via curvature-aware contrastive learning, projecting multimodal features into a unified manifold space that explicitly preserves geometric and semantic relationships. Subsequently, CGI leverages manifold curvature as a geometric cue to refine cross-modal attention, enabling fine-grained multimodal alignment by explicitly associating high-curvature visual and textual features.

In this study, we propose ControlFuse, a controllable IVIF framework that achieves multi-granularity fusion guided by user instructions. As illustrated in Fig. 1(c), this approach supports effective fusion control at global, semantic, and instance granularities, resulting in enhanced IVIF adaptability for practical applications. The contributions are summarized as follows:

- We introduce ControlFuse, a controllable IVIF framework guided by user instructions at global, semantic, and instance levels. To the best of our knowledge, it is the first to support such multi-granularity guidance, enabling precise and flexible adaptation of the fusion behavior to application requirements.
- We construct an IVIF dataset via an automated annotation pipeline that explicitly aligns textual instructions with spatial masks at multiple levels of granularity. This multi-granularity dataset addresses the scarcity of spatially annotated multimodal data for IVIF, enabling precise and flexible instruction-guided fusion learning under multi-level supervision.
- We design a Multimodal Feature Interaction Module (MFIM) comprising Feature Manifold Converter (FMC) and Curvature-Guided Interaction (CGI) to achieve fine-grained alignment between textual instructions and visual features. FMC maps both modalities into a unified manifold space, while CGI exploits manifold curvature to guide cross-modal attention. Trained with curvature-aware contrastive learning, this module facilitates precise instruction-guided IVIF.

Related Work

Controllable Image Fusion

Recently, user instructions have been integrated into various computer vision tasks, enabling flexible control over image synthesis (He et al. 2025), semantic decision-making (Zang et al. 2025), and related applications. In the context of image fusion, researchers have also started exploring instruction-guided frameworks that leverage textual descriptions as auxiliary inputs to achieve flexible and controllable fusion outcomes. Text-IF (Yi et al. 2024) uses textual inputs to restore degraded scenes (low light, noise, low contrast), but its adjustments remain global rather than fine-grained. TextFusion (Cheng et al. 2025) advances text-guided fusion by associating textual semantics with visual features, enabling semantic-level control, yet fusion granularity is still limited to category-level guidance without precise instance-level enhancement. OmniFuse (Zhang et al. 2025) introduces language-driven semantic control in a latent diffusion framework to handle complex degradations, but, similar to TextFusion, operates only at the semantic level and lacks explicit instance-level control, limiting its effectiveness in scenarios requiring precise object-focused enhancement. The latest TIRN (Wang et al. 2025b) achieves highlights for user-defined instances, but is limited to instance-level control. In contrast, our proposed framework, ControlFuse, addresses the critical gap in existing methods by explicitly supporting multi-granularity fusion control, from global and semantic

down to precise instance-level guidance. This capability significantly expands the flexibility and practical applicability of instruction-guided IVIF, enabling users to tailor fusion outcomes according to task-specific requirements.

Manifold Geometric Properties

The manifold view in computer vision (Wang et al. 2025c; Park 2005) stems from the observation that high-dimensional data (e.g., images and text) lie on lower-dimensional, nonlinear manifolds in feature space. Such representations compactly encode intrinsic geometry through properties like curvature and distance. In multimodal tasks, modeling image and text data as manifolds naturally facilitates structured cross-modal alignment while preserving modality-specific geometric information.

Manifold curvature highlights structurally salient image regions (edges, contours) and semantically distinctive text tokens, providing natural anchors for fine-grained multimodal alignment and attention. Manifold distances further quantify feature proximity within and across modalities, enabling adaptive matching between visual and textual features. Together, these geometric properties improve precise, context-aware alignment, directly benefiting controllable IVIF.

Methodology

Problem Formulation

Conventional image fusion methods take the visible and infrared images (I_{ir}, I_{vis}) as inputs to a fusion network ζ_n, producing a fixed fused image. Formally, this fusion process approximates a predefined fusion function $\mathcal{F}_{if}$:

$$I_f = \mathcal{F}_{if}\left(I_{vis}, I_{ir}; \zeta_n\right). \tag{1}$$

Because the learned fusion strategy is fixed, these methods cannot adapt to specific fusion preferences according to user requirements.

In contrast, the recent text-guided fusion paradigm introduces textual inputs to achieve more flexible fusion control, enabling global or semantic-level adjustments. In this paradigm, the fusion task is reformulated as:

$$I_f = \mathcal{F}_{if}^{t}\left(I_{vis}, I_{ir}, T; \zeta_n^{t}\right), \tag{2}$$

where the original fusion function $\mathcal{F}_{if}$ is extended to $\mathcal{F}_{if}^{t}$ under textual guidance T. By incorporating text semantics, the fusion network ζ_n^{t} provides controllability beyond fixed outcomes. However, it is inadequate for addressing instance-specific fusion requirements, i.e., differentiating among individual objects within the same semantic category.

We extend the above fusion paradigm to integrate multi-granularity textual instructions, enabling precise fusion control at global, semantic, and instance levels. Formally, our refined paradigm is expressed as:

$$I_f = \mathcal{F}_{if}^{mgt}\left(I_{vis}, I_{ir}, T; \zeta_n^{mgt}\right). \tag{3}$$

The fusion function $\mathcal{F}_{if}^{t}$ is extended to $\mathcal{F}_{if}^{mgt}$ by integrating multi-granularity textual instructions T at global, semantic, and instance levels. Through the proposed multimodal feature interaction module, comprising Feature Manifold Converter (FMC) and Curvature-Guided Interaction (CGI), the fusion network ζ_n^{mgt} adaptively aligns textual instructions with corresponding visual features. Consequently, our network achieves multi-granularity fusion control, supporting user instructions ranging from global scene and semantic-level to instance-specific enhancement.

Image Fusion Pipeline

As illustrated in Fig. 2, the image fusion pipeline consists of three stages: multimodal feature encoding, multimodal feature interaction, and reconstruction.

Multimodal Feature Encoding. The first stage extracts high-level features from both the input images and user-provided instructions. We adopt Restormer-based blocks (Zhao et al. 2023b) to encode the visible image $I_{vis} \in R^{H\times W\times 3}$ and the infrared image $I_{ir} \in R^{H\times W\times 1}$:

$$\Phi_{vis} = E_v^I\left(I_{vis}\right), \quad \Phi_{ir} = E_i^I\left(I_{ir}\right), \tag{4}$$

where $\Phi_{vis} \in \mathbb{R}^{C_{vis}\times H\times W}$ and $\Phi_{ir} \in \mathbb{R}^{C_{ir}\times H\times W}$ denote the encoded visible and infrared features, respectively. E_v^I and E_i^I represent their corresponding encoders. To capture textual semantic information, we employ BLIP-2 (Li et al. 2023), a pretrained vision-language model known for its robust linguistic representation. Its parameters remain frozen to preserve pretrained linguistic consistency. Formally, the textual embedding extraction with frozen parameters $\{\cdot\}_p$ is defined as:

$$\Phi_t = \left\{E_t^T\right\}_p(T) \in \mathbb{R}^{d_t}, \tag{5}$$

where T denotes the textual description, d_t indicates the embedding dimension.

Multimodal Feature interaction. After extracting the visual features $\Phi_{\text{vis}}, \Phi_{\text{ir}}$ and the textual features Φ_t, we project them into a unified manifold using FMC. FMC employs a learnable transformation $\mathcal{F}$ to embed each modality into a common d-dimensional space:

$$\Phi'_{\text{vis}} = \mathcal{F}(\Phi_{\text{vis}}), \quad \Phi'_{\text{ir}} = \mathcal{F}(\Phi_{\text{ir}}), \quad \Phi'_t = \mathcal{F}(\Phi_t), \tag{6}$$

where $\Phi'_{\text{vis}}, \Phi'_{\text{ir}} \in \mathbb{R}^{d\times H\times W}$ and $\Phi'_t \in \mathbb{R}^{d\times 1\times 1}$. By mapping features onto this unified manifold, FMC prepares modalities for effective cross-modal interaction.

Next, we employ CGI $\mathcal{E}(\cdot)$ to generate object-specific location maps $\{\mathcal{M}_k\}$ guided by textual instructions. CGI leverages curvature signals within the visual features (Φ'_{vis}, Φ'_{ir}) and matches them with textual features Φ'_t to highlight spatial regions corresponding to each object O_k specified by the user:

$$\mathcal{M}_k = \mathcal{E}\big(\Phi'_{\text{vis}},\, \Phi'_{\text{ir}},\, \Phi'_t\big). \tag{7}$$

To effectively utilize $\mathcal{M}_k$, we concatenate visual features along the channel dimension: $\Phi_V = \big[\Phi'_{\text{vis}},\ \Phi'_{\text{ir}}\big]$. Then, we spatially modulate Φ_V with the location map $\mathcal{M}_k$, projected via a 1×1 convolution to match channel dimensions: $\Phi_{\mathcal{M}} = \text{Conv}_{1\times 1}(\mathcal{M}_k)$. Finally, we perform cross-attention $\mathcal{C}(\cdot)$ by flattening Φ_V into query tokens Q_V and using $\Phi_{\mathcal{M}}$ as the corresponding key-value representations K_M, V_M:

$$\Phi_F \ = \ \mathcal{C}\big(Q_V,\, K_M,\, V_M\big). \tag{8}$$

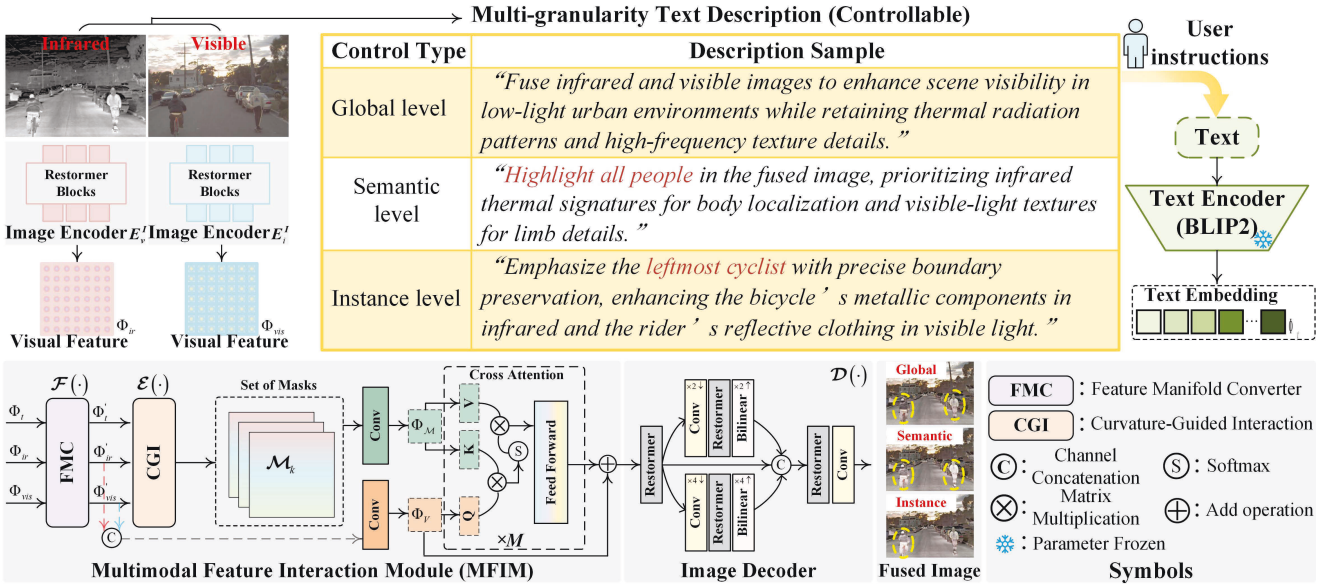

Figure 2: Overview of the proposed multi-granularity controllable fusion framework.

This attention mechanism allows visual tokens to adaptively focus on regions indicated by the generated location maps. A skip connection further preserves essential visual information: $\Phi_F \leftarrow \Phi_F + \Phi_V$.

Reconstruction. Finally, the fused features Φ_F are input into a multi-scale decoder $\mathcal{D}$, composed of N Restormer blocks and a 1×1 convolution for channel projection. Through iterative refinement and multi-scale feature upsampling, $\mathcal{D}$ generates the fused image I_F, formally written as $I_F = \mathcal{D}(\Phi_F)$.

Multimodal Feature Interaction Module

The MFIM enables precise cross-modal alignment. It consists of two components:

Feature Manifold Converter. FMC projects visual and textual features into a unified manifold space that preserves geometric and semantic relationships. We first represent visual features $\mathbf{\Phi}_{vis}$, $\mathbf{\Phi}_{ir} \in \mathbb{R}^{C\times H\times W}$ as flattened spatial sequences $\mathbb{R}^{HW\times C}$ and textual features $\mathbf{\Phi}_t \in \mathbb{R}^{d_t\times L}$ as sequential token embeddings. These embeddings form the nodes of a multimodal graph $G = (V, E)$. Edges are constructed based on intra-modal adjacency (spatial for visual, sequential for textual) and cross-modal feature similarity. Through L-layer Graph Attention Networks (GAT), each node aggregates intra- and cross-modal information, producing aligned manifold features: $\mathbf{h}_v^{(l+1)} \leftarrow \text{GAT}(\mathbf{h}_v^{(l)}, \mathcal{N}(v))$. Finally, linear projections map the node embeddings back to structured feature spaces, yielding unified manifold features $\mathbf{\Phi}'_{vis}$, $\mathbf{\Phi}'_{ir} \in \mathbb{R}^{d\times H\times W}$, and $\mathbf{\Phi}'_t \in \mathbb{R}^{d\times L}$.

Curvature-Guided Interaction. CGI leverages structural curvature as geometric guidance for fine-grained cross-modal alignment. We first compute spatial curvature maps for visual features ($\mathbf{\Phi}'_{vis}$, $\mathbf{\Phi}'_{ir}$) and sequential curvature for textual embeddings ($\mathbf{\Phi}'_t$). The visual curvature κ_{vis}, κ_{ir} highlights salient structures (edges or object boundaries), computed via second-order spatial derivatives per feature channel:

$$\kappa_{vis} = \|\nabla^2 \mathbf{\Phi}'_{vis}\|_2, \quad \kappa_{ir} = \|\nabla^2 \mathbf{\Phi}'_{ir}\|_2. \tag{9}$$

Text curvature κ_t captures distinct semantic transitions between consecutive textual tokens:

$$\kappa_t(l) = \|\mathbf{\Phi}'_t(l-1) - 2\mathbf{\Phi}'_t(l) + \mathbf{\Phi}'_t(l+1)\|_2. \tag{10}$$

Next, we concatenate visual features along the channel dimension to form queries $Q_V \in \mathbb{R}^{HW\times 2d}$, while textual embeddings serve as keys $K_t \in \mathbb{R}^{L\times d}$. Curvature values modulate the cross-attention weights via a learnable parameter ω:

$$\mathbf{A} = \text{softmax}\left(\frac{Q_V K_t^T}{\sqrt{d}} + \omega(\kappa_V + \kappa_t)\right), \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{HW\times L}$. For each object O_k mentioned in the text, we aggregate attention scores from corresponding textual tokens to form a spatial location map: $\mathbf{M}_k = \sum_{l\in\mathcal{I}_k} \mathbf{A}[:, l]$. The location map guide spatial fusion, precisely aligning user-specified textual instructions with corresponding visual regions.

Loss Function

Our controllable IVIF framework employs a multi-term loss, jointly optimizing object localization for precise spatial guidance, content fidelity for structural consistency, and curvature-based alignment for robust cross-modal correspondence. **Object Localization Loss ($\mathcal{L}_m$).** We predict spatial location maps $\mathcal{M}_k$ guided by textual instructions and optimize their accuracy using the binary cross-entropy loss: $\mathcal{L}_m = -\sum_{k,h,w} M_k^{gt}(h,w) \log\left(\mathcal{M}_k(h,w)\right)$, where M_k^{gt} denotes the ground-truth spatial mask explicitly associated with the k-th textual instruction. **Content Fidelity Loss ($\mathcal{L}_f$).** We enforce content fidelity using pixel-level (l_1) and

structural similarity (SSIM) losses. The pixel-level loss is $\mathcal{L}_{pix} = \|I_f - I_{ir}\|_1 + \|I_f - I_{vis}\|_1$. The structural similarity loss is defined as $\mathcal{L}_{ssim} = 2 - \text{SSIM}(I_f, I_{ir}) - \text{SSIM}(I_f, I_{vis})$. The total content fidelity loss combines both terms simply as $\mathcal{L}_f = \mathcal{L}_{pix} + \mathcal{L}_{ssim}$. **Curvature-based Alignment Loss ($\mathcal{L}_a$).** We embed IR, VIS, and textual features into a unified manifold space, assigning each embedding a curvature value $(\kappa_{ir}(i), \kappa_{vis}(j), \kappa_{text}(l))$ indicating structural significance. For each IR-VIS-Text triplet positive (matched) (i, j, l) and negative (mismatched) $(i, j, \tilde{l})$, the overall alignment loss is defined by contrasting positive ($\mathcal{P}$) and negative ($\mathcal{N}$) textual tokens:

$$\mathcal{L}_a = \sum_{(i,j,l)\in\mathcal{P}} w(i,j,l)\, d(i,j,l) - \sum_{(i,j,\tilde{l})\in\mathcal{N}} w(i,j,\tilde{l})\, d(i,j,\tilde{l}), \tag{12}$$

where the alignment distance is $d(i,j,l) = \|\Psi_{ir}(i) - \Psi_{vis}(j)\|_2 + \|\Psi_{ir}(i) - \Psi_{text}(l)\|_2 + \|\Psi_{vis}(j) - \Psi_{text}(l)\|_2$, and the curvature-based weight is $w(i,j,l) = 1 + \lambda[\kappa_{ir}(i) + \kappa_{vis}(j) + \kappa_{text}(l)]$. Minimizing $\mathcal{L}_a$ enforces alignment between matched multimodal embeddings, emphasizing structurally salient visual regions and semantically distinctive text tokens. **Total Loss.** We adopt uncertainty-based adaptive weighting (Kendall, Gal, and Cipolla 2018) to dynamically balance our multi-objective loss terms. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_m^2}\mathcal{L}_m + \frac{1}{2\sigma_f^2}\mathcal{L}_f + \frac{1}{2\sigma_a^2}\mathcal{L}_a + \log(\sigma_m\sigma_f\sigma_a), \tag{13}$$

where $\sigma_m, \sigma_f, \sigma_a$ are learnable uncertainty parameters jointly optimized during training. This strategy allows stable and automatic balancing among $\mathcal{L}_m$, $\mathcal{L}_f$, and $\mathcal{L}_a$.

Dataset

Automatic mask generation. Following the training splits in Datasets and Metrics, i.e., 171 / 3,780 / 1,083 VIS–IR pairs from RoadScene, M3FD, and MSRS (5,034 pairs in total), we build a multi-granularity IVIF training set. For each pair, category-prompted object detectors (Liu et al. 2024c) are applied to both VIS and IR images, their proposals are filtered and merged, and the fused boxes are fed into SAM (Kirillov et al. 2023) to obtain instance-level masks. Instance masks are grouped into semantic masks by category, while the whole image domain is treated as the global mask. We further keep at most K salient instances per image. The overall pipeline is summarized in Algorithm 1.

Instruction construction. For each image pair, we generate natural-language instructions for the global, semantic, and selected instance masks. Global instructions describe scene-level fusion goals, semantic instructions specify category-wise emphasis or suppression, and instance instructions refer to individual salient objects with coarse spatial cues. This yields multi-granularity instruction–mask pairs aligned with the three control levels in our framework.

Negative instructions for contrastive learning. On top of positive instruction–mask pairs, we synthesize negative instructions that intentionally mismatch the masks by referring to absent categories or incorrect spatial relations. During training, these negative instructions are stored without ground-truth masks and used only as hard negative text tokens in the curvature-aware alignment loss, reducing similarity for visually inconsistent descriptions while preserving high similarity for correctly matched pairs.

Algorithm 1: Construction of Multi-Granularity Dataset

Input: VIS-IR pairs $\mathcal{D}_{\text{train}}$; category prompts $\mathcal{C}$
Output: Instruction-mask set $\mathcal{S}$

$\mathcal{S} \leftarrow \varnothing$
for all $(I^{\text{vis}}, I^{\text{ir}}) \in \mathcal{D}_{\text{train}}$ **do**
  $(B^{\text{vis}}_{\text{raw}}, s^{\text{vis}}) \leftarrow \text{GroundingDINO}(I^{\text{vis}}, \mathcal{C})$ // VIS-side boxes & scores
  $(B^{\text{ir}}_{\text{raw}}, s^{\text{ir}}) \leftarrow \text{IRDetector}(I^{\text{ir}}, \mathcal{C})$ // IR-side boxes & scores
  $B^{\text{vis}} \leftarrow \text{FilterBoxes}(B^{\text{vis}}_{\text{raw}}, s^{\text{vis}}, \tau^{\text{vis}}_{\text{score}}, \tau_{\text{area}})$
  $B^{\text{ir}} \leftarrow \text{FilterBoxes}(B^{\text{ir}}_{\text{raw}}, s^{\text{ir}}, \tau^{\text{ir}}_{\text{score}}, \tau_{\text{area}})$
  $B \leftarrow \text{MergeBoxes}(B^{\text{vis}}, B^{\text{ir}}, \tau_{\text{iou}})$ // union of VIS-only, IR-only and common objects
  $M^{\text{ins}} \leftarrow \text{SAM}(I^{\text{vis}}, I^{\text{ir}}, B)$ // instance-level masks
  Group M^{ins} by category to obtain semantic masks $\{M^{\text{sem}}_c\}$
  $M^{\text{glob}} \leftarrow$ whole image domain // global mask
  $\mathcal{K}(I) \leftarrow \text{SelectSalientInstances}(M^{\text{ins}}, K)$ // select up to K salient instances
  $M^{\text{ins}}_{\mathcal{K}(I)} \leftarrow \{M^{\text{ins}}_k\}_{k\in\mathcal{K}(I)}$
  $T^{\text{glob}} \leftarrow \text{GenGlobalInst}(M^{\text{glob}})$
  $T^{\text{sem}} \leftarrow \{\text{GenSemanticInst}(c, M^{\text{sem}}_c)\}_c$
  $T^{\text{ins}} \leftarrow \{\text{GenInstanceInst}(k, M^{\text{ins}}_k)\}_{k\in\mathcal{K}(I)}$
  $\tilde{T} \leftarrow \text{GenNegativeInst}\begin{pmatrix} T^{\text{glob}}, T^{\text{sem}}, T^{\text{ins}}, \\ M^{\text{sem}}, M^{\text{ins}} \end{pmatrix}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \left\{ \begin{matrix} (T^{\text{glob}}, M^{\text{glob}}),\ (T^{\text{sem}}, M^{\text{sem}}), \\ (T^{\text{ins}}, M^{\text{ins}}_{\mathcal{K}(I)}),\ (\tilde{T}, \varnothing) \end{matrix} \right\}$
end for
return $\mathcal{S}$

Experiment

Training Details. We train the network for 100 epochs using the AdamW optimizer, with an initial learning rate of 1×10^{-4} and decreasing by a factor of 0.5 every 20 epochs. The batch size is set to 8. Experiments are conducted on the NVIDIA GeForce RTX 4090 GPU with PyTorch.

Datasets and Metrics. We verified the fusion performance on three datasets: RoadScene(Xu et al. 2020), M3FD(Liu et al. 2022), and MSRS(Tang et al. 2022). The RoadScene dataset contains 221 pairs, split into 171 for training and 50 for testing. M^3FD provides 4,204 pairs (3,780 train and 424 test), and MSRS provides 1,444 pairs (1,083 train and 361 test). Fusion quality is evaluated using six metrics: EN, SD, SF, AG, VIF, and Qabf.

Comparison with SOTA Methods

We evaluate ControlFuse against various state-of-the-art methods including EMMA (Zhao et al. 2024a), Text-IF (Yi et al. 2024), TextFusion (Cheng et al. 2025), MTG-Fusion (Wang et al. 2025c), OmniFuse (Zhang et al. 2025), DCEvo

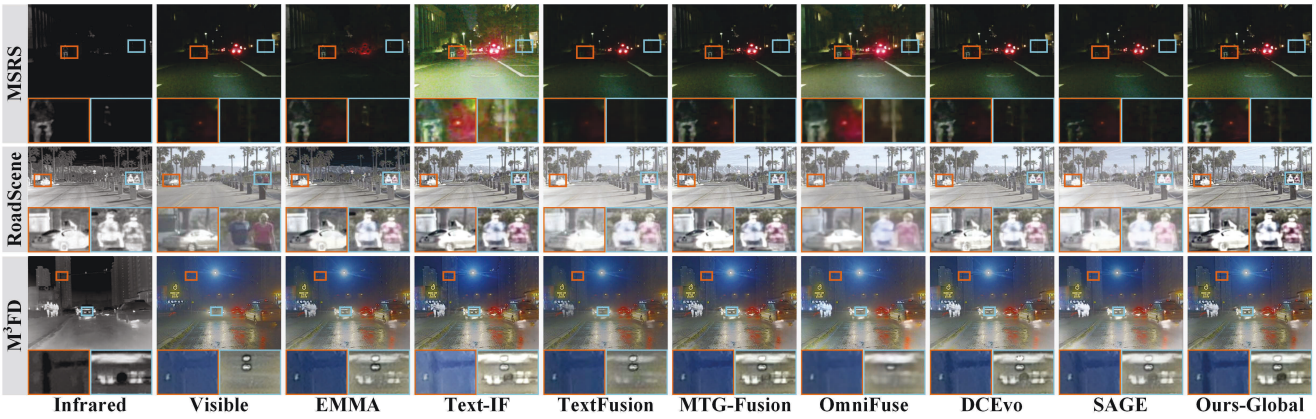

Figure 3: Qualitative comparisons of our ControlFuse (global-level) and existing image fusion methods. From top to bottom: low-light in MSRS, low-quality in RoadScene, and high-brightness in M^{3}FD.

Methods [Pub./Year]	Dataset: MSRS						Dataset: RoadScene						Dataset: M^{3}FD					
	EN	SD	SF	AG	VIF	Qabf	EN	SD	SF	AG	VIF	Qabf	EN	SD	SF	AG	VIF	Qabf
EMMA [CVPR'24]	5.85	54.48	11.34	3.68	1.01	0.71	4.80	53.54	12.12	4.66	0.69	0.51	6.10	59.92	15.88	6.21	0.88	0.74
Text-IF [CVPR'24]	5.91	59.61	11.24	3.66	1.02	0.71	4.94	52.09	12.53	4.61	0.71	0.52	6.11	64.21	15.53	6.30	0.88	0.74
TextFusion [InF'25]	5.87	56.30	11.31	3.64	1.06	0.73	4.96	54.61	12.42	4.55	0.73	0.53	6.13	63.58	15.62	6.03	0.85	0.74
MTG-Fusion [IJCV'25]	5.98	**61.44**	11.60	4.05	**1.08**	0.72	5.05	55.35	**13.18**	4.90	0.81	**0.59**	6.21	67.89	**16.68**	6.57	0.91	0.78
OmniFuse [TPAMI'25]	5.79	53.74	11.23	3.46	1.00	0.69	4.79	51.71	11.94	4.51	0.69	0.47	6.12	58.02	15.35	6.18	0.88	0.73
DCEvo [CVPR'25]	5.90	60.26	11.58	3.91	1.06	0.72	4.98	52.46	12.74	4.78	0.75	0.53	6.15	63.30	16.09	6.39	0.90	0.76
SAGE [CVPR'25]	5.91	60.56	11.80	3.96	**1.08**	0.75	5.06	54.71	12.70	4.73	0.77	0.57	6.17	65.03	16.46	6.37	0.89	0.79
ControlFuse-Global	**5.99**	61.15	**11.94**	**4.15**	1.07	**0.76**	**5.19**	**55.96**	13.06	**5.05**	**0.85**	**0.59**	**6.26**	**68.05**	16.60	**6.82**	**0.92**	**0.81**

Table 1: Quantitative comparison of fusion performance with other methods. **Bold**: best; Underline: second-best.

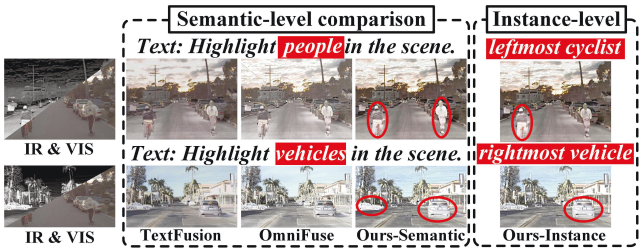

Figure 4: Verification of ControlFuse's controllability. Further comparison of the semantic-level instruction control results and display of instance-level instruction control results.

(Liu et al. 2025a), and SAGE (Wu et al. 2025).

Qualitative Comparison. For fair comparison, ControlFuse adopts global-level textual instructions. As shown in Figure 3, ControlFuse effectively fuses thermal cues and visible textures, yielding clearer targets in dark regions, stronger foreground–background separation, and sharper structures than competing methods.

Quantitative Comparison. Table 1 reports results on the MSRS, RoadScene, and M^{3}FD datasets. Under global textual instructions, ControlFuse achieves superior performance on most metrics, indicating robust fusion quality across diverse illumination conditions and target categories.

Controllability comparison. We further assess controllability in Fig. 4. With semantic-level instructions, ControlFuse enhances the specified category and better suppresses irrelevant regions than TextFusion and OmniFuse; with instance-level instructions, it precisely highlights the target instance with clear boundaries, demonstrating effective multi-granularity control.

Performance on High-level Tasks

We evaluate ControlFuse on downstream tasks and verify its controllability using multi-granularity instructions.

Object Detection. Experiments on M^{3}FD with an 8:1:1 train/val/test split use YOLOv12 (Tian, Ye, and Doermann 2025) as the detector. As shown in Fig. 5(a), ControlFuse detects all people while other methods miss targets, and semantic/instance-level instructions further improve confidence over global instructions; quantitatively, Table 2(a) shows ControlFuse achieves the highest mAP@[.5], validating the benefit of multi-granularity control for detection.

Semantic Segmentation. Semantic segmentation is evaluated on MSRS using RTFNet (Sun, Zuo, and Liu 2019) as backbone. ControlFuse exploits global, semantic, and instance-level instructions (e.g., "Enhance the entire scene," "Highlight all people," "Highlight the CarStop sign"). As shown in Fig. 5(b), it yields more accurate masks, and Table 2(b) shows consistently higher segmentation accuracy, validating the benefits of multi-granularity control.

Methods	(a) Detection on M^{3}FD Dataset							(b) Segmentation on MSRS Dataset							
	People	Car	Bus	Motorcycle	Lamp	Truck	@.5	Car	Person	Bike	Curve	CarStop	Cone	Bump	mIoU
IR	0.339	0.450	0.031	0.549	0.017	0.142	0.113	0.605	0.688	0.295	0.193	0.238	0.160	0.171	0.374
VIS	0.524	0.631	0.705	0.477	0.603	0.515	0.470	0.677	0.680	0.467	0.269	0.358	0.234	0.220	0.428
EMMA	0.579	0.758	0.729	0.550	0.616	0.546	0.561	0.807	0.705	0.49	0.313	0.455	0.264	0.222	0.469
Text-IF	0.615	0.741	0.759	0.582	0.628	0.567	0.574	0.825	0.713	0.486	0.318	0.457	0.267	0.230	0.476
TextFusion	0.622	0.762	0.764	0.557	0.642	0.585	0.589	0.832	0.725	0.490	0.327	0.468	0.268	0.232	0.476
MTG-Fusion	0.700	0.817	0.764	0.590	0.672	0.623	0.612	0.844	0.721	0.509	0.339	0.479	0.271	**0.240**	0.493
OmniFuse	0.608	0.731	0.722	0.536	0.652	0.563	0.600	0.814	0.720	0.501	0.317	0.471	0.262	0.231	0.476
DCEvo	0.640	0.793	0.768	0.596	0.681	0.583	0.597	0.842	0.719	0.513	0.328	0.474	0.268	0.236	0.487
SAGE	0.686	0.794	0.762	0.600	0.657	0.593	0.619	0.832	0.723	0.506	0.333	0.483	0.267	0.233	0.482
Ours-Global	0.715	0.794	0.807	0.612	0.685	0.640	0.621	0.843	0.727	0.510	0.339	0.480	0.275	0.237	0.489
Ours-Semantic	**0.773**	**0.832**	**0.841**	**0.647**	**0.707**	**0.689**	**0.666**	**0.865**	**0.744**	**0.514**	**0.340**	**0.485**	**0.278**	0.233	**0.496**

Table 2: Quantitative comparison results on downstream tasks, i.e., object detection and semantic segmentation.

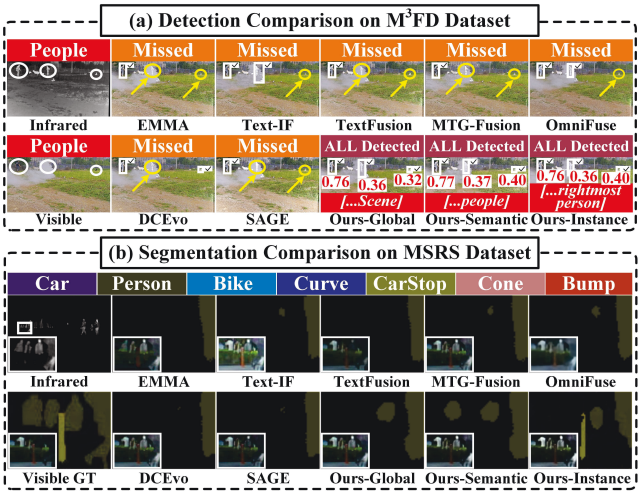

Figure 5: Downstream object detection and semantic segmentation on fused images. With multi-granularity instructions, our method achieves better downstream performance.

Ablation Study

Effect of Multimodal Feature Interaction Module. Table 3(a) shows that removing FMC and keeping only CGI degrades all metrics, indicating CGI alone cannot maintain cross-modal alignment. Keeping FMC but removing CGI partially recovers performance, underscoring the role of manifold-preserving feature conversion. The baseline without both modules performs worst, confirming that FMC and CGI are complementary and jointly critical for accurate localization and perceptual quality.

Effect of Multi-Term Loss. Table 3(b) shows that removing $\mathcal{L}_m$ harms localization, discarding $\mathcal{L}_f$ reduces structural fidelity (SF), and omitting $\mathcal{L}_a$ weakens cross-modal alignment with more semantic mismatches, while using all three losses yields the best performance.

Effect of Negative Prompt. We assess negative prompts in curvature-aware contrastive learning. As shown in Fig. 6, removing negative prompts weakens discrimination between relevant and irrelevant instructions, lowering segmentation IoU and F1, whereas including them improves text-object alignment and yields higher IoU/F1 over training.

Variant	Setting	EN	SD	SF	AG	VIF	Qabf
(a) Multimodal Feature Interaction Module							
CGI-only	FMC$^-$/CGI$^+$	4.63	36.83	9.72	3.12	0.67	0.55
FMC-only	FMC$^+$/CGI$^-$	4.78	39.15	10.51	3.42	0.69	0.59
Baseline	FMC$^-$/CGI$^-$	4.52	34.22	8.93	2.87	0.62	0.51
(b) Loss Terms							
w/o $\mathcal{L}_m$	$\mathcal{L}_f$+$\mathcal{L}_a$	4.81	40.21	9.14	2.25	0.62	0.57
w/o $\mathcal{L}_a$	$\mathcal{L}_m$+$\mathcal{L}_f$	4.73	39.82	9.85	2.35	0.60	0.54
w/o $\mathcal{L}_f$	$\mathcal{L}_m$+$\mathcal{L}_a$	3.65	37.43	8.37	2.02	0.46	0.45
Full		**5.99**	**61.15**	**11.94**	**4.15**	**1.07**	**0.76**

Table 3: Ablation study of multimodal feature interaction module and loss terms on the MSRS dataset.

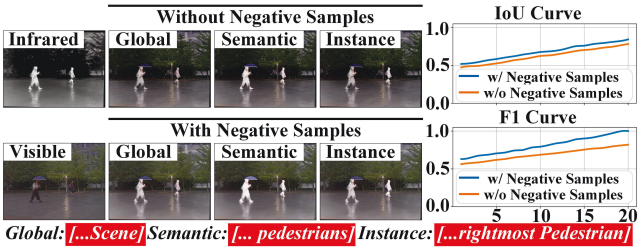

Figure 6: Ablation on negative prompt generation. Including negative prompts in curvature-aware contrastive learning improves instruction-guided text-region alignment compared with using positive prompts only.

Conclusion

This study presents a controllable IVIF framework that extends text-guided fusion from global and semantic control to multi-granularity control, enabling instance-level and instruction-consistent fusion. We introduce a multimodal interaction module that combines manifold conversion and curvature-guided cross-modal interaction to align instructions with relevant image regions. We also build a VIS-IR dataset with instruction–masking pairs at global, semantic, and instance levels to support training. Experiments on standard benchmarks and downstream tasks show consistent improvements over conventional fusion and prior text-guided methods, especially for instance-level instructions. Future work will focus on more robust instruction-region supervision and better generalization to complex scenes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [No. 62401097]; 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

Cheng, C.; Xu, T.; Wu, X.-J.; Li, H.; Li, X.; Tang, Z.; and Kittler, J. 2025. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117: 102790.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2025. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17123–17131.

Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, 539–555. Springer.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

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.; Liu, J.; Zhang, Y.; and Liu, Y. 2024. A deep learning framework for infrared and visible image fusion without strict registration. *International Journal of Computer Vision*, 132(5): 1625–1644.

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025a. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; Wang, D.; Hu, Z.; Li, S.; Ni, W.; Zhao, L.; and Wang, Q. 2025b. Fd2-net: Frequency-driven feature decomposition network for infrared-visible object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4797–4805.

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*, 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.; and Fan, X. 2024b. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, J.; Zhang, B.; Mei, Q.; Li, X.; Zou, Y.; Jiang, Z.; Ma, L.; Liu, R.; and Fan, X. 2025a. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2226–2235.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, Y.; Zou, Y.; Li, X.; Zhu, X.; Han, K.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Toward a Training-Free Plug-and-Play Refinement Framework for Infrared and Visible Image Registration and Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1268–1277.

Park, J. 2005. *Manifold learning in computer vision*. The Pennsylvania State University.

Sun, Y.; Zuo, W.; and Liu, M. 2019. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3): 2576–2583.

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*.

Tang, L.; Wang, Y.; Cai, Z.; Jiang, J.; and Ma, J. 2025a. ControlFusion: A Controllable Image Fusion Framework with Language-Vision Degradation Prompts. *arXiv preprint arXiv:2503.23356*.

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*, 1–19.

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: 79–92.

Tang, W.; He, F.; and Liu, Y. 2022. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25: 5413–5428.

Tang, W.; He, F.; and Liu, Y. 2024. ITFuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition*, 156: 110822.

Tian, Y.; Ye, Q.; and Doermann, D. 2025. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.

Wang, Z.; Zhang, J.; Guan, T.; Zhou, Y.; Li, X.; Dong, M.; and Liu, J. 2025a. Efficient Rectified Flow for Image Fusion. *Advances in Neural Information Processing Systems*.

Wang, Z.; Zhang, J.; Song, H.; Ge, M.; Wang, J.; and Duan, H. 2025b. Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12637–12647.

Wang, Z.; Zhao, L.; Zhang, J.; Song, R.; Song, H.; Meng, J.; and Wang, S. 2025c. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 1–23.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17882–17891.

Xu, H.; Ma, J.; Le, Z.; Jiang, J.; and Guo, X. 2020. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12484–12491.

Yang, Z.; Zhang, Y.; Li, H.; and Liu, Y. 2025. Instruction-driven fusion of Infrared–visible images: Tailoring for diverse downstream tasks. *Information Fusion*, 121: 103148.

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.

Zang, Y.; Li, W.; Han, J.; Zhou, K.; and Loy, C. C. 2025. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2): 825–843.

Zhang, H.; Cao, L.; Zuo, X.; Shao, Z.; and Ma, J. 2025. OmniFuse: Composite Degradation-Robust Image Fusion with Language-Driven Semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Zhang, X.; Ye, P.; and Xiao, G. 2020. VIFB: A visible and infrared image fusion benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 104–105.

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13955–13965.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023b. 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.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024a. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023c. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8082–8093.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; et al. 2024b. Image fusion via vision-language model. *arXiv preprint arXiv:2402.02235*.