TRIREWARD-FUSION: UNIFIED VISION-LANGUAGE REWARD LEARNING FOR TASK-ORIENTED INFRARED–VISIBLE IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared and visible image fusion aims to combine the thermal saliency of infrared images with the structural details of visible images for robust high-level visual perception. However, most existing fusion methods are optimized using handcrafted no-reference pixel-level objectives or supervised by a single downstream task, making it difficult to jointly preserve visual fidelity, semantic integrity, and object-level reliability. Moreover, image fusion, semantic segmentation, and object detection pursue distinct and sometimes conflicting objectives, and a simple weighted combination of their losses often fails to provide unified and semantically meaningful feedback. To address this problem, we propose TriReward-Fusion, a BLIP-guided vision-language reward learning framework for task-oriented infrared-visible image fusion. The proposed framework is trained on MSRS by jointly using image-pair no-reference fusion objectives, semantic segmentation annotations, and our additionally annotated detection labels. It projects fusion, segmentation, and detection outputs into a shared vision-language semantic state space, where perceptual information preservation, semantic region completeness, and instance-level detection reliability are transformed into learnable multi-task reward signals. These rewards provide unified semantic feedback to guide the fusion process and reconcile pixel-level reconstruction, region-level parsing, and object-level localization objectives. Experiments evaluate fusion quality on MSRS, M3FD, and LLVIP, semantic segmentation on MSRS, and object detection on M3FD. The results show that TriReward-Fusion improves fused image quality and enhances downstream segmentation and detection performance, demonstrating the effectiveness of vision-language reward learning for unified multi-task infrared-visible image fusion.

1 INTRODUCTION

Infrared and visible imaging exhibit inherent complementarity. Infrared images can highlight salient targets such as pedestrians and vehicles according to thermal radiation differences between objects and the background, providing strong target perception capability under nighttime, low-light, and adverse-weather conditions. Visible images, in contrast, contain rich textures, contours, and structural details, which are beneficial for scene understanding and object recognition Ma et al. (2019a). Infrared-visible image fusion (IVF) aims to integrate complementary information from the two modalities into a single image, thereby providing a more complete, stable, and discriminative visual representation for intelligent surveillance, autonomous driving, and complex scene perception Hwang et al. (2015); DeSouza & Kak (2002). In recent years, fusion methods based on convolutional neural networks (CNN) Ram Prabhakar et al. (2017), Transformers Xu et al. (2020), generative adversarial networks (GAN) Ma et al. (2019b), and diffusion models Saharia et al. (2022) have achieved substantial progress in cross-modal feature extraction, information interaction, and image reconstruction. However, generating fused images that simultaneously exhibit high visual quality and consistently improve high-level visual perception remains a key challenge in IVF.

Because ideal fused images usually lack ground-truth (GT) references, most IVF methods are optimized with handcrafted no-reference objectives, such as intensity, gradient, structural similarity, and information entropy Zhang et al. (2020). Although these low-level constraints help preserve infrared

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

saliency and visible texture details, they mainly rely on pixel statistics and local structural consistency, and thus cannot reliably assess whether the fused image retains semantics useful for high-level scene understanding Bai et al. (2025). Consequently, fused images with high contrast, strong edges, or large information content may still suffer from weakened thermal targets, damaged object boundaries, over-enhanced backgrounds, or incomplete semantic regions, indicating that low-level visual objectives alone are insufficient for task-oriented IVF.

To bridge this gap, recent studies have introduced downstream perception tasks, such as semantic segmentation and object detection, into fusion training Tang et al. (2022a); Liu et al. (2022a). However, most of them focus on a single task and do not fully exploit the complementary region-level and instance-level feedback provided by segmentation and detection. Moreover, fusion, segmentation, and detection have different optimization preferences, and simply combining their losses may cause gradient competition and objective bias. Vision-language models (VLMs), pretrained on large-scale image-text pairs, offer a potential shared semantic reference for evaluating fusion quality and downstream perception reliability Yi et al. (2024); Cheng et al. (2025). Nevertheless, directly using generic VLM scores is unstable because of prompt sensitivity Schmalfuss et al. (2025), domain discrepancy, and limited adaptation to fine-grained boundaries Moshtaghi et al. (2025), small objects, and cross-modal saliency in IVF Peng et al. (2024). Therefore, this work focuses on transforming VLM semantic knowledge into task-specific, learnable reward signals that can coordinate fusion, segmentation, and detection within a unified optimization framework.

Based on the above analysis, we propose TriReward-Fusion, a unified vision-language reward learning framework for task-driven IVF. Using BLIP Li et al. (2022; 2023b) as a semantic prior, it projects the fused image, segmentation prediction, and detection prediction into a shared vision-language semantic state space, where perceptual preservation, semantic region completeness, and instance-level detection reliability are modeled as three learnable rewards. These rewards are adaptively aggregated with conventional fusion, segmentation, and detection objectives to provide fusion-oriented semantic feedback. In addition, a segmentation-detection consistency constraint coordinates semantic masks and bounding boxes, alleviating multi-task conflicts and improving both fusion quality and downstream perception performance.

To support joint training and evaluation of the three tasks under aligned image pairs, we additionally annotate instance-level bounding boxes for the person, bicycle, and car categories on the MSRS Tang et al. (2022b) dataset, which originally provides infrared-visible image pairs and semantic segmentation labels. It should be noted that this detection-label extension is mainly used to construct a unified multi-task training environment, while the core focus of this work remains vision-language reward learning and multi-task collaborative optimization. In the experiments, we evaluate fusion quality on MSRS, M3FD Liu et al. (2022b), and LLVIP Jia et al. (2021), semantic segmentation performance on MSRS, and object detection performance on M3FD, in order to validate the effectiveness and generalization ability of the proposed method under different datasets, scenarios, and task settings. The main contributions of this work are summarized as follows:

- We propose TriReward-Fusion, a unified framework that jointly optimizes IVF, semantic segmentation, and object detection for task-driven fusion.
- We design a BLIP-guided semantic state reward mechanism that learns three task-specific rewards for perceptual preservation, semantic completeness, and detection reliability.
- We develop a multi-reward collaborative optimization strategy with adaptive reward aggregation and segmentation-detection consistency to reduce task conflicts and improve downstream perception.

2 Related work

IVF has been extensively studied to integrate thermal saliency from infrared images and texture details from visible images. Recent methods have gradually evolved from low-level feature fusion toward semantic-aware and foundation-model-guided fusion. PIAFusion Tang et al. (2022c) introduced illumination-aware progressive fusion to improve the adaptability of infrared-visible fusion under different lighting conditions. CDDFuse Zhao et al. (2023) decomposed cross-modal features into shared and modality-specific components through a correlation-driven dual-branch architecture. IGNet Li et al. (2023a) employed a graph neural network to model cross-modal interactions between

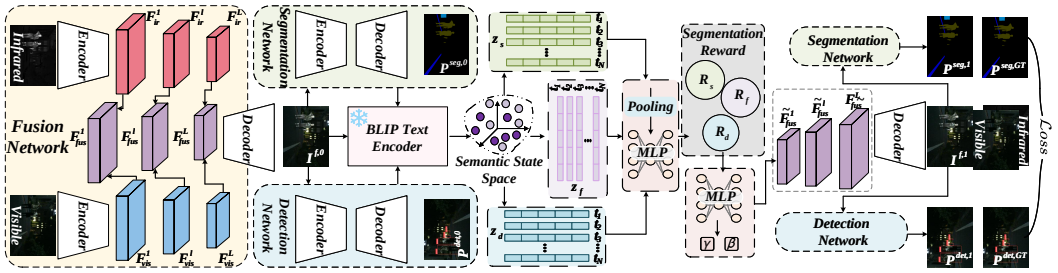


Figure 1: Overall architecture of the proposed TriReward-Fusion framework. It maps fusion, segmentation, and detection outputs into a BLIP-guided semantic space, and uses learned tri-level rewards to refine the fused image and downstream predictions.

infrared and visible images, providing a graph-based perspective for image fusion. Dif-Fusion Yue et al. (2023) introduced diffusion models into infrared-visible fusion to improve color fidelity and generative quality. PFCFuse Hu et al. (2024) combined PoolFormer and CNN architectures to balance global context modeling and local detail preservation. SHIP Zheng et al. (2024) investigated synergistic high-order interactions in infrared-visible fusion to enhance cross-modal information modeling. IF-FILM Zhao et al. (2024) incorporated vision-language models into image fusion, enabling the fusion process to exploit high-level semantic priors. Text-DiFuse Zhang et al. (2024) developed a text-modulated diffusion framework for interactive multi-modal image fusion. SAGE-SAM Wu et al. (2025) introduced SAM-based semantic priors to provide region-level guidance for multi-modality image fusion. MaeFuse Li et al. (2025) transferred pretrained masked autoencoder features to improve representation learning and training stability in infrared-visible fusion. PTS-GAN Song et al. (2025) incorporated text-semantic awareness into a progressive GAN framework to improve semantic consistency and visual quality in image fusion. OmniFuse Zhang et al. (2025) used language-driven semantics to enhance robustness under composite degradation scenarios.

3 METHOD

3.1 UNIFIED MULTI-TASK FUSION FRAMEWORK

Given a spatially aligned infrared-visible image pair $I^{ir} \in \mathbb{R}^{1\times H\times W}$ and $I^{vis} \in \mathbb{R}^{3\times H\times W}$, TriReward-Fusion learns a unified model $\mathcal{F}_\Theta$ that predicts a fused image, a semantic segmentation map, and object detection results:

$$(I^f, P^{seg}, P^{det}) = \mathcal{F}_\Theta(I^{ir}, I^{vis}). \tag{1}$$

Rather than treating the three tasks as independent objectives, the framework maps their outputs into a BLIP-guided semantic state space and derives three task-specific rewards, R_f, R_s, and R_d. These rewards measure perceptual information preservation, semantic region completeness, and instance-level detection reliability, respectively, and provide task-aware feedback for optimizing fusion,as shown in Figure 1.

We first extract multi-scale features using two modality-specific encoders:

$$\mathbf{F}_m^l = \mathcal{E}_m^l(\mathbf{F}_m^{l-1}), \quad m \in \{ir, vis\},\ l = 1, \dots, L, \tag{2}$$

where $\mathbf{F}_{ir}^0 = I^{ir}$ and $\mathbf{F}_{vis}^0 = I^{vis}$. At each scale, infrared and visible features are channel-aligned and fused by a cross-modal interaction module:

$$\mathbf{F}_{fus}^l = \mathcal{M}^l\big[\widehat{\mathbf{F}}_{ir}^l, \widehat{\mathbf{F}}_{vis}^l, |\widehat{\mathbf{F}}_{ir}^l - \widehat{\mathbf{F}}_{vis}^l|, \widehat{\mathbf{F}}_{ir}^l \odot \widehat{\mathbf{F}}_{vis}^l\big], \tag{3}$$

where $[\cdot]$ denotes channel concatenation and $\odot$ denotes element-wise multiplication. The fusion decoder then reconstructs the initial fused image $I^{f,0} = \mathcal{D}_f(\{\mathbf{F}_{fus}^l\}_{l=1}^L)$. To preserve task-specific representations, segmentation and detection use separate encoder-decoder branches:

$$P^{seg,0} = \mathcal{D}_{seg}\left(\mathcal{E}_{seg}(I^{f,0}, \{\mathbf{F}_{fus}^l\}_{l=1}^L)\right), \quad P^{det,0} = \mathcal{D}_{det}\left(\mathcal{E}_{det}(I^{f,0}, \{\mathbf{F}_{fus}^l\}_{l=1}^L)\right). \tag{4}$$

The initial outputs are evaluated by a task-supervised vision-language reward model. It consists of frozen BLIP visual and text encoders and three trainable reward heads, parameterized by Ω_f, Ω_s, and Ω_d. To avoid co-adaptation between the reward model and the task network, the reward heads are first calibrated with task-derived quality targets while the task network is fixed. The calibrated reward model is then frozen and used as a differentiable critic.

3.2 BLIP-GUIDED VISION-LANGUAGE SEMANTIC STATE CONSTRUCTION

Directly using a generic vision-language score for image fusion is unstable because the score depends on prompts, image order, and the domain gap between natural RGB images and infrared-visible imagery. We therefore use BLIP to build a shared semantic state space, rather than using its raw matching score as a fusion-quality objective. Let $\mathcal{B}_{vis}$ and $\mathcal{B}_{text}$ denote the frozen BLIP visual and text encoders. For each category c shared by semantic segmentation and object detection, we construct a prompt set $\mathcal{T}_c = \{T_c^m\}_{m=1}^M$. The category text prototype is obtained by prompt ensembling:

$$\mathbf{t}_c = \mathrm{Norm}\left[\frac{1}{M}\sum_{m=1}^{M}\mathcal{P}_t\left(\mathcal{B}_{text}(T_c^m)\right)\right], \quad c \in \mathcal{C}_{sd}, \tag{5}$$

where $\mathcal{P}_t$ projects BLIP text features into a d-dimensional semantic space, and $\mathcal{C}_{sd}$ is the category set shared by segmentation and detection. For the infrared, visible, and fused images, the BLIP visual tokens are

$$\mathbf{V}_m = \mathrm{Norm}\left[\mathcal{P}_v\left(\mathcal{B}_{vis}(\widetilde{I}^m)\right)\right] \in \mathbb{R}^{N_v \times d}, \quad m \in \{ir, vis, f\}, \tag{6}$$

where $\widetilde{I}^m$ denotes the three-channel BLIP input. For infrared images, the intensity channel is repeated three times. To associate visual regions with category semantics, we compute a text-guided attention weight $a_{m,c,i}$ from the similarity between visual token $\mathbf{V}_{m,i}$ and text prototype $\mathbf{t}_c$, with temperature τ_v. The category-conditioned image state is $\mathbf{z}_{m,c} = \sum_i a_{m,c,i}\mathbf{V}_{m,i}$, and the global image state is $\overline{\mathbf{z}}_m = \mathrm{Pool}(\mathbf{V}_m)$. For semantic segmentation, the intermediate task features and soft probability map are projected into the same space:

$$\mathbf{U}_s = \mathrm{Norm}\left[\mathcal{P}_s\left(\phi_s(\mathbf{Z}_{seg}, P^{seg})\right)\right] \in \mathbb{R}^{N_s \times d}. \tag{7}$$

For detection, the encoder-decoder features are combined with the continuous class, objectness, and box predictions before non-maximum suppression:

$$\mathbf{U}_d = \mathrm{Norm}\left[\mathcal{P}_d\left(\phi_d(\mathbf{Z}_{det}, P_{cls}, P_{obj}, P_{box})\right)\right] \in \mathbb{R}^{N_d \times d}. \tag{8}$$

The category-specific segmentation and detection states are extracted by text-guided cross-attention:

$$\mathbf{z}_{s,c} = \mathrm{CrossAttn}(\mathbf{t}_c, \mathbf{U}_s, \mathbf{U}_s), \quad \mathbf{z}_{d,c} = \mathrm{CrossAttn}(\mathbf{t}_c, \mathbf{U}_d, \mathbf{U}_d). \tag{9}$$

The reward descriptors are constructed from these semantic states. The fusion descriptor aggregates $\mathbf{z}_{ir,c}$, $\mathbf{z}_{vis,c}$, $\mathbf{z}_{f,c}$, and the two source-fusion differences over all shared categories, together with the global states. The segmentation descriptor aggregates $\mathbf{z}_{f,c}$, $\mathbf{z}_{s,c}$, $\mathbf{t}_c$, and their semantic differences. The detection descriptor is defined analogously using $\mathbf{z}_{d,c}$. The reward scores are produced by lightweight heads:

$$R_f = \mathcal{H}_f([\overline{\mathbf{z}}_{ir}, \overline{\mathbf{z}}_{vis}, \overline{\mathbf{z}}_f, \mathbf{g}_f]), \quad R_s = \mathcal{H}_s(\mathbf{g}_s), \quad R_d = \mathcal{H}_d(\mathbf{g}_d). \tag{10}$$

This design makes BLIP category prototypes participate in all reward branches, rather than serving as isolated text embeddings. The reward heads share the same output range but do not share parameters, so each branch can model task-specific quality while remaining comparable after reward normalization.

3.3 TASK-SUPERVISED TRI-REWARD CALIBRATION

Directly maximizing jointly learned reward scores can cause degenerate co-adaptation, where the reward heads increase their outputs without improving task quality. We therefore calibrate the three reward branches with task-derived supervision before using them to optimize the fusion network. Because no GT fused image is available, the fusion reward is learned by relative preference. Given an initial fused image $I^{f,+}$, we construct a degraded counterpart $I^{f,-} = \mathcal{D}_q(I^{f,+})$. The degradation operator randomly applies infrared-target attenuation, visible-texture suppression, boundary blurring, local over-exposure, or contrast degradation. These perturbations reduce at least one desired fusion property while preserving scene content. The ranking loss is

$$\mathcal{L}_{rank}^f = -\log\sigma\left(\frac{R_f^+ - R_f^- - m}{\tau_r}\right), \tag{11}$$

where m is a margin and τ_r is a temperature parameter. For semantic segmentation, the reward target combines region overlap and boundary preservation. The soft class-wise IoU is

$$Q^s_{iou} = \frac{1}{|\mathcal{C}_{seg}|} \sum_{c \in \mathcal{C}_{seg}} \frac{A_c}{B_c}, A_c = \sum_{x,y} P^{seg}_c(x,y) Y^{seg}_c(x,y), B_c = \sum_{x,y} P^{seg}_c(x,y) + \sum_{x,y} Y^{seg}_c(x,y) - A_c + \epsilon, \tag{12}$$

where Y^{seg} is the ground-truth segmentation. Let Q^s_{bd} be the boundary F-score between predicted and ground-truth boundaries. The segmentation target and calibration loss are

$$R^*_s = \alpha_s Q^s_{iou} + (1 - \alpha_s) Q^s_{bd}, \quad \mathcal{L}^s_{reg} = \mathrm{SmoothL1}(R_s, R^*_s). \tag{13}$$

For object detection, continuous predictions are matched to ground-truth instances by bipartite matching. For each matched pair $(j,k) \in \mathcal{M}$, the instance quality is

$$q^d_{j,k} = \alpha_{box} \,\mathrm{IoU}(\widehat{\mathbf{b}}_j, \mathbf{b}^{gt}_k) + \alpha_{cls} P^{j,y_k}_{cls} + \alpha_{obj} P^j_{obj}, \quad \alpha_{box} + \alpha_{cls} + \alpha_{obj} = 1. \tag{14}$$

The image-level detection target averages matched instance quality over $\max(N_{gt}, 1)$ and subtracts a false-positive penalty proportional to $N_{fp}/(N_{pred} + \epsilon)$. It is clipped to $[0,1]$ to stabilize reward regression. The detection calibration loss is $\mathcal{L}^d_{reg} = \mathrm{SmoothL1}(R_d, R^*_d)$.

We further align task states with BLIP category prototypes. The category-presence label $y_c \in \{0,1\}$ is derived from segmentation and detection annotations. For each $\xi \in \{f, s, d\}$, the compatibility score $q_{\xi,c}$ is computed from the cosine similarity between $\mathbf{z}_{\xi,c}$ and $\mathbf{t}_c$ with temperature τ_a. The vision-language alignment loss is the average BCE over all shared categories and the three task states. The complete reward-model calibration objective is

$$\mathcal{L}_{RM} = \lambda^r_f \mathcal{L}^f_{rank} + \lambda^r_s \mathcal{L}^s_{reg} + \lambda^r_d \mathcal{L}^d_{reg} + \lambda_{vl} \mathcal{L}_{vl}. \tag{15}$$

Let Θ denote the task-network parameters, Φ the frozen BLIP parameters, and Ω the projection layers and reward heads. During calibration, $\Theta^{(0)}$ and Φ are fixed, and only Ω is optimized:

$$\Omega^* = \arg\min_{\Omega} \mathcal{L}_{RM}(\Theta^{(0)}, \Phi, \Omega). \tag{16}$$

3.4 Reward-Guided Collaborative Optimization

After calibration, the BLIP encoders and reward model Ω^* are frozen. The rewards serve as fixed differentiable critics: their parameters are not updated, but gradients from reward outputs to fused images and downstream predictions are retained. Because R_f, R_s, and R_d have different numerical distributions, each reward is standardized by running statistics:

$$\widehat{R}_\tau = \frac{R_\tau - \mu_\tau}{\sigma_\tau + \epsilon}, \quad \tau \in \{f, s, d\}. \tag{17}$$

The reward weights are predicted from the deepest fused feature:

$$\boldsymbol{\omega} = \frac{\rho}{3}\mathbf{1} + (1 - \rho)\,\mathrm{Softmax}\left[\mathcal{G}\left(\mathrm{GAP}(\mathbf{F}^L_{fus})\right)\right]. \tag{18}$$

This gives $\sum_\tau \omega_\tau = 1$ and $\omega_\tau \geq \rho/3$. The reward-guidance loss is

$$\mathcal{L}_{guide} = -\sum_{\tau \in \{f,s,d\}} \omega_\tau \widehat{R}_\tau(\mathbf{x}_\tau; \Omega^*), \tag{19}$$

where $\mathbf{x}_\tau$ denotes the corresponding task state. During this stage, $\nabla_{\Omega^*} \mathcal{L}_{guide} = 0$ and $\nabla_\Theta \mathcal{L}_{guide} \neq 0$, so the task network is optimized through a fixed quality evaluator. To provide feature-level feedback, the normalized rewards are converted into a quality state $\mathbf{r} = \mathrm{sg}([\widehat{R}_f, \widehat{R}_s, \widehat{R}_d])$ and embedded by MLP_r. For scale l, the embedding predicts affine parameters $(\boldsymbol{\gamma}^l, \boldsymbol{\beta}^l)$, and the fused feature is modulated as

$$\widetilde{\mathbf{F}}^l_{fus} = (1 + \boldsymbol{\gamma}^l) \odot \mathbf{F}^l_{fus} + \boldsymbol{\beta}^l. \tag{20}$$

The fusion decoder reconstructs the refined fused image $I^{f,1}$ from $\{\widetilde{\mathbf{F}}^l_{fus}\}^L_{l=1}$. The same task-specific branches then produce $P^{seg,1}$ and $P^{det,1}$ from the refined image and modulated features.

The decoder weights are shared between the initial and reward-refined stages, so reward feedback does not introduce another set of decoders.

The task network is trained with both the initial and refined predictions. For stage $k \in \{0, 1\}$, the supervised objective is

$$\mathcal{L}_{sup}^{k} = \mathcal{L}_{fus}^{k} + \mathcal{L}_{seg}^{k} + \mathcal{L}_{det}^{k}. \tag{21}$$

The fusion loss combines intensity, gradient, and structural-similarity terms. The segmentation loss combines cross-entropy and Dice losses, and the detection loss follows the standard classification, box regression, and distribution focal losses. The total objective is

$$\mathcal{L}_{total} = \mathcal{L}_{sup}^{1} + \eta \mathcal{L}_{sup}^{0} + \lambda_r \mathcal{L}_{guide} + \lambda_c \mathcal{L}_{cross}. \tag{22}$$

where η controls auxiliary supervision on the initial outputs, while λ_r and λ_c balance reward guidance and cross-task consistency. Finally, we impose segmentation-detection consistency on refined predictions. For detection candidate j with class c_j and box $\mathbf{b}_j = (x_j^1, y_j^1, x_j^2, y_j^2)$, we construct a differentiable soft rectangular mask

$$G_j(x, y) = \sigma[\kappa(x - x_j^1)]\sigma[\kappa(x_j^2 - x)]\sigma[\kappa(y - y_j^1)]\sigma[\kappa(y_j^2 - y)]. \tag{23}$$

The average segmentation response inside this box is

$$s_j = \frac{\sum_{x,y} G_j(x, y) P_{c_j}^{seg,1}(x, y)}{\sum_{x,y} G_j(x, y) + \epsilon}. \tag{24}$$

The confidence-weighted box-mask loss is

$$\mathcal{L}_{bm} = \frac{\sum_{j=1}^{N} w_j (1 - s_j)}{\sum_{j=1}^{N} w_j + \epsilon}, \quad w_j = P_{obj}^{j} P_{cls}^{j,c_j}. \tag{25}$$

To avoid numerical saturation from multiplying many probabilities, category presence is estimated by top-K pooling. p_{seg}^{c} is computed from the top-K_s pixels of $P_c^{seg,1}$, and p_{det}^{c} from the top-K_d detection confidences $\{P_{obj}^{j} P_{cls}^{j,c}\}_{j=1}^{N_d}$. Their discrepancy is measured by the Jensen-Shannon divergence between Bernoulli distributions. The final cross-task consistency loss is

$$\mathcal{L}_{cross} = \mathcal{L}_{bm} + \lambda_{pc} \frac{1}{|\mathcal{C}_{sd}|} \sum_{c \in \mathcal{C}_{sd}} \mathrm{JS}_{Ber}(p_{seg}^{c} \| p_{det}^{c}). \tag{26}$$

4 Experiment

4.1 Setup

4.1.1 Datasets

Experiments are conducted on three infrared-visible datasets: MSRS, M3FD, and LLVIP. MSRS is used as the only dataset for training TriReward-Fusion. It provides spatially aligned infrared-visible image pairs and pixel-level semantic segmentation labels. To support unified three-task training, we additionally annotate bounding boxes for three instance-level categories, including person, bicycle, and car. The annotations are produced by jointly inspecting infrared and visible images and are manually checked to reduce missing or incorrect boxes caused by small targets, occlusion, and adjacent instances. We follow the original MSRS split Tang et al. (2022c) and further separate a validation set from the training set for hyperparameter selection and early stopping. The same split is used for fusion, segmentation, and detection to avoid data leakage. M3FD and LLVIP are used only for cross-dataset evaluation. No images or labels from M3FD or LLVIP are used to train, calibrate, or fine-tune TriReward-Fusion. After training on MSRS, the learned fusion model is directly applied to M3FD and LLVIP to generate fused images. LLVIP is used only for fusion-quality evaluation. This protocol ensures that the comparison on M3FD and LLVIP measures the generalization of the generated fused images rather than additional task-specific fine-tuning of the proposed model.

Table 1: Quantitative comparison results of different fusion methods on different datasets. Red and blue denote the best and second-best results.

Datasets	Method	Source	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM
MSRS	PIAFusion	22'InfFus	6.64	45.34	12.12	2.75	1.70	1.02	0.67	1.00
	CDDFuse	23'CVPR	6.12	33.91	9.66	1.29	1.56	0.69	0.47	0.67
	IGNet	23'MM	6.70	43.38	11.56	3.47	1.62	1.05	0.69	1.01
	Dif-fusion	23'TIP	6.54	42.44	11.89	2.25	1.61	0.82	0.58	0.84
	PFCFuse	24'TMM	6.69	43.12	11.60	2.65	1.63	0.90	0.61	0.91
	SHIP	24'CVPR	6.43	41.13	11.81	2.86	1.51	0.91	0.66	0.90
	IF-FILM	24'ICML	6.68	44.05	12.15	2.78	1.41	0.82	0.62	0.90
	Text-DiFuse	24'NeurIPS	6.71	44.26	12.37	2.89	1.45	0.84	0.69	0.62
	SAGE-SAM	25'CVPR	6.00	36.34	10.42	2.23	1.42	0.71	0.54	0.88
	MaeFuse	25'TIP	6.84	46.28	12.19	2.04	1.81	1.05	0.68	0.92
	PTS-GAN	25'PR	6.70	44.15	12.25	2.82	1.43	0.83	0.64	0.81
	OmniFuse	25'TPAMI	6.75	45.10	12.20	3.52	1.60	0.92	0.63	0.80
	Ours	–	6.95	47.12	12.65	3.42	1.78	1.12	0.72	0.98
M3FD	PIAFusion	22'InfFus	6.87	34.01	12.49	2.60	1.43	0.94	0.60	0.96
	CDDFuse	23'CVPR	6.94	39.93	12.17	1.45	1.54	0.74	0.45	0.74
	IGNet	23'MM	6.85	37.47	12.87	2.27	1.60	0.90	0.60	1.02
	Dif-fusion	23'TIP	6.59	32.71	11.41	2.01	1.19	0.48	0.42	0.78
	PFCFuse	24'TMM	6.85	37.62	13.09	2.15	1.55	0.69	0.51	0.76
	SHIP	24'CVPR	6.63	32.46	13.12	2.91	1.11	0.86	0.62	0.88
	IF-FILM	24'ICML	6.88	35.10	12.60	2.45	1.50	0.72	0.56	0.75
	Text-DiFuse	24'NeurIPS	6.85	37.90	13.09	2.69	1.28	0.94	0.61	0.76
	SAGE-SAM	25'CVPR	6.85	36.46	12.73	2.16	1.69	0.72	0.58	0.97
	MaeFuse	25'TIP	7.40	49.56	20.72	1.90	1.83	0.83	0.57	0.82
	PTS-GAN	25'PR	6.90	38.20	18.95	2.50	1.65	0.75	0.66	0.88
	OmniFuse	25'TPAMI	7.02	41.50	14.25	2.55	1.62	0.85	0.71	0.90
	Ours	–	7.31	45.20	21.15	2.85	1.78	0.96	0.73	0.99
LLVIP	PIAFusion	22'InfFus	7.03	42.89	12.24	8.60	1.53	0.80	0.52	0.75
	CDDFuse	23'CVPR	7.01	42.46	10.95	8.38	1.54	0.80	0.49	0.76
	IGNet	23'MM	6.90	42.99	10.68	8.41	1.52	0.74	0.41	0.59
	Dif-fusion	23'TIP	7.07	43.34	11.11	8.63	1.47	0.71	0.44	0.69
	PFCFuse	24'TMM	7.01	41.43	11.74	8.38	1.54	0.78	0.51	0.76
	SHIP	24'CVPR	7.02	42.54	12.18	8.69	1.41	0.79	0.52	0.71
	IF-FILM	24'ICML	7.10	44.50	8.50	8.42	1.45	0.72	0.51	0.62
	Text-DiFuse	24'NeurIPS	7.34	49.65	8.93	8.58	1.70	0.67	0.33	0.46
	SAGE-SAM	25'CVPR	6.34	39.01	10.54	8.05	1.26	0.61	0.43	0.64
	MaeFuse	25'TIP	7.43	58.95	15.61	8.57	1.72	0.87	0.49	0.66
	PTS-GAN	25'PR	7.35	45.80	8.70	8.48	1.50	0.94	0.47	0.84
	OmniFuse	25'TPAMI	7.48	48.10	13.05	8.72	1.61	0.82	0.50	0.72
	Ours	–	7.58	50.40	16.85	8.66	1.76	0.95	0.58	0.86

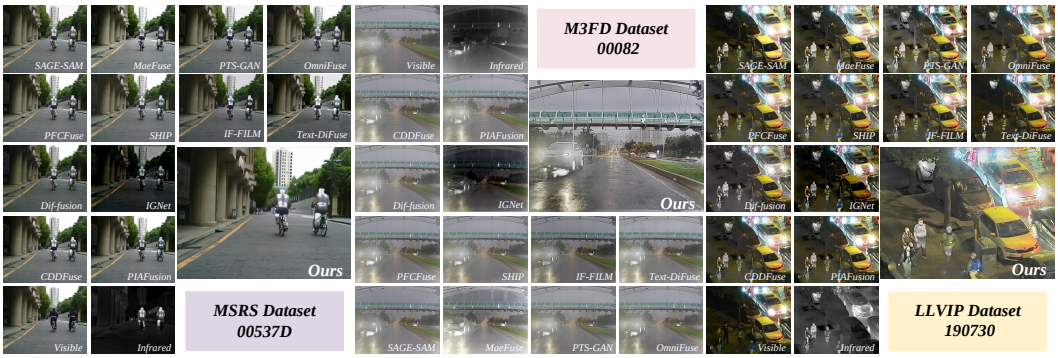


Figure 2: Qualitative comparison results of different fusion methods on different datasets.

4.1.2 METRICS

For the quantitative analysis of the fusion results, we employ a comprehensive set of eight established evaluation metrics, Entropy (EN) Shannon (1948), Standard Deviation (SD) and Spatial Frequency (SF) Eskicioglu & Fisher (1995), Mutual Information (MI) Qu et al. (2002), Sum of Correlations of Differences (SCD) Aslantas & Bendes (2015), Visual Information Fidelity (VIF) Sheikh

Table 2: Ablation study results of TriReward-Fusion.

Configuration	EN	VIF	$Q^{AB/F}$	mIoU	mAP
w/o Tri-reward	5.84	1.04	0.65	74.15	81.82
w/o Adaptive aggregation	6.12	1.06	0.67	74.82	82.35
w/o Reward modulation	6.35	1.07	0.68	75.14	82.90
w/o Cross-task consistency	6.68	1.09	0.70	75.60	83.45
Full Model	**6.95**	**1.12**	**0.72**	**76.32**	**84.11**

& Bovik (2006), $\mathbf{Q^{AB/F}}$ Xydeas & Petrovic (2000), and Structural Similarity (SSIM) Wang et al. (2004). Higher metric values correspond to better fused image quality.

4.1.3 Implementation Details

All experiments are implemented in PyTorch on a single NVIDIA GeForce RTX 2080 Ti GPU. We use AdamW with a weight decay of 1×10^{-4}, initial learning rates of 1×10^{-4} for task-network parameters and 1×10^{-5} for the pretrained visual backbone, and cosine annealing with a 5-epoch linear warm-up. The batch size is 4, with automatic mixed precision enabled. Training follows the two-stage procedure described in the Method section. During the first 40 epochs, only conventional fusion, segmentation, and detection losses are used to obtain stable initial predictions. The reward heads are then calibrated using task-derived quality targets while the task network and BLIP encoders remain fixed. During the remaining 80 epochs, the calibrated reward model is frozen for reward-guided optimization, with reward-conditioned feature modulation and segmentation-detection consistency enabled. The reward weight λ_r is linearly increased from 0 to its target value over the first 10 epochs of this stage.

4.1.4 Comparison Protocol

For fair comparison, all methods, including TriReward-Fusion, are reproduced using identical data splits and experimental settings. Each method generates fused training, validation, and test sets. We assess downstream task applicability using SegFormer-B3 Xie et al. (2021) for semantic segmentation and YOLOv8m Jocher et al. (2023) for object detection. For each method, downstream models are independently trained on its fused training set and evaluated on the corresponding test set under identical architectures, pretrained weights, input resolutions, augmentations, optimization settings, training epochs, and model-selection criteria. Test labels are used only for final evaluation; full reproduction details are provided in the supplementary material.

4.2 Comparison with Existing Methods

We test TriReward-Fusion on the three test sets and compare the fusion results with the existing methods including PIAFusion Tang et al. (2022c), CDDFuse Zhao et al. (2023), IGNet Li et al. (2023a), Dif-fusion Yue et al. (2023), PFCFuse Hu et al. (2024), SHIP Zheng et al. (2024), IF-FILM Zhao et al. (2024), Text-DiFuse Zhang et al. (2024), SAGE-SAM Wu et al. (2025), MaeFuse Li et al. (2025), PTS-GAN Song et al. (2025) and OmniFuse Zhang et al. (2025).

As shown in Figure 2, compared with other methods, our approach better preserves infrared thermal salient targets while maintaining visible textures and structural details across MSRS, M3FD, and LLVIP. The results also show fewer artifacts and less over-smoothing, demonstrating the effectiveness of TriReward-Fusion in generating perception-friendly fused images.

As shown in Table 1, we evaluate our method quantitatively on three datasets.

4.3 Ablation Study

To evaluate the contribution of each component in TriReward-Fusion, we conduct component-wise ablation experiments, as reported in Table 2. Removing the tri-reward mechanism leads to the largest degradation across all metrics, showing that BLIP-guided rewards provide useful semantic supervision beyond conventional fusion and task losses. The drops in EN, VIF, and $Q^{AB/F}$ indicate weaker information preservation and structural fidelity, while the decreases in mIoU and mAP suggest reduced suitability for downstream perception.

Table 3: Quantitative comparison of different fusion methods on downstream semantic segmentation and object detection tasks. Red and blue indicate the best and second-best results.

Methods	MSRS		M3FD	
	mIoU	**mAcc**	**mAP**$_{50}$	**mAP**$_{50:95}$
Visible	69.34	75.12	71.87	46.87
Infrared	45.20	52.40	55.77	37.10
PIAFusion	74.14	78.20	81.46	55.87
CDDFuse	74.21	78.35	81.80	56.21
IGNet	70.15	76.05	79.63	54.62
Dif-fusion	71.83	77.10	82.27	55.57
PFCFuse	64.69	71.20	82.69	56.39
SHIP	73.72	77.95	81.06	55.50
IF-FILM	74.19	78.30	81.58	56.02
Text-DiFuse	72.64	77.50	73.74	46.03
SAGE-SAM	72.95	78.79	82.46	57.26
MaeFuse	74.16	78.25	80.99	54.95
PTS-GAN	73.95	78.10	81.22	55.63
OmniFuse	74.58	79.12	82.85	57.30
Ours	76.32	79.89	84.11	58.42

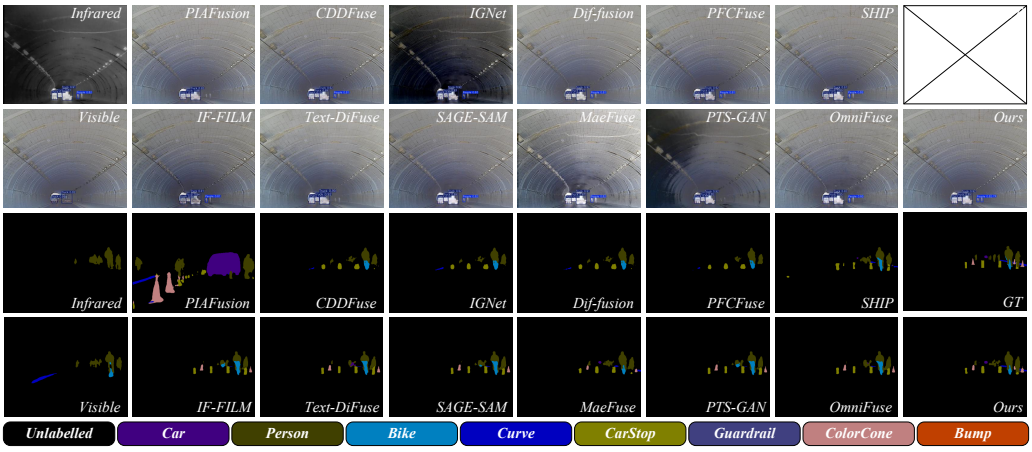


Figure 3: Qualitative comparison results of different fusion methods on downstream applications.

Adaptive reward aggregation and reward-conditioned feature modulation also bring consistent gains. This indicates that dynamically balancing R_f, R_s, and R_d, and injecting reward states into multi-scale features, help coordinate pixel-level reconstruction, region-level parsing, and instance-level localization. Removing cross-task consistency further reduces mIoU and mAP, confirming that segmentation and detection offer complementary region- and instance-level cues.

4.4 Downstream Task Experiment

As shown in Figure 3, we visualize the downstream perception results produced. These results show that the fused images produced by our framework preserve task-relevant semantic information and provide reliable support for downstream perception. As shown in Table 3, we quantitatively evaluate the downstream task performance of different fusion methods.

5 Conclusion

In this paper, we proposed TriReward-Fusion, a unified vision-language reward learning framework for task-driven infrared-visible image fusion. The framework integrates image fusion, semantic segmentation, and object detection, and uses BLIP-guided semantic rewards to provide perceptual, region-level, and instance-level feedback. By combining adaptive reward aggregation, reward-conditioned feature modulation, and segmentation-detection consistency, TriReward-Fusion improves fusion quality and downstream perception performance across diverse infrared-visible scenarios. These results demonstrate the effectiveness of vision-language reward learning for task-oriented fusion optimization and unified perception.

REFERENCES

Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of Electronics and Communications*, 69 (12):1890–1896, 2015.

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7457–7468, 2025.

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. doi: 10.1016/j.inffus.2024.102790.

Guilherme N. DeSouza and Avinash C. Kak. Vision for mobile robot navigation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):237–267, 2002. doi: 10.1109/34.982903.

Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995.

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.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1037–1045, 2015. doi: 10.1109/CVPR.2015.7298706.

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*, pp. 3496–3504, 2021.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL `https://github.com/ultralytics/ultralytics`.

Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4471–4479, 2023a.

Jiayang Li, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. Maefuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. *IEEE Transactions on Image Processing*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pp. 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742, 2023b.

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*, pp. 5802–5811, 2022a.

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*, pp. 5802–5811, 2022b.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019a.

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, 2019b.

Mehdi Moshtaghi, Siavash H. Khajavi, and Joni Pajarinen. Rgb-th-bench: A dense benchmark for visual-thermal understanding of vision language models. *arXiv preprint arXiv:2503.19654*, 2025.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13279–13288, 2024.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313–315, 2002.

K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4714–4722, 2017.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andres Bruhn, and Jose M. Alvarez. Parc: A quantitative framework uncovering the symmetries within vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25081–25091, 2025.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

Wenhao Song, Mingliang Gao, Qilei Li, Gwanggil Jeon, and David Camacho. Progressive text-semantic-aware generative adversarial network for image fusion. *Pattern Recognition*, pp. 112720, 2025.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022a. doi: 10.1016/j.inffus.2021.12.004.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022b. ISSN 1566-2535.

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, 2022c.

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.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17882–17891, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2020.

Costas S Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

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*, pp. 27026–27035, 2024.

Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32:5705–5720, 2023.

Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12797–12804, 2020. doi: 10.1609/aaai.v34i07.6975.

Hao Zhang, Lei Cao, and Jaiyi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

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.

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*, pp. 5906–5916, 2023.

Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *Proceedings of the International Conference on Machine Learning*, 2024.

Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26384–26395, 2024.

A Appendix

A.1 Additional Dataset Details

To enable joint three-task training on MSRS, we manually annotate instance-level bounding boxes for three categories: person, bicycle, and car. Two annotators independently label all images by inspecting both infrared and visible modalities. Afterwards, cross-checking is performed to filter false-positive boxes and recover missing annotations introduced by small targets, occlusion, and adjacent instances. These manually annotated detection labels on MSRS will be publicly released together with our source code upon paper acceptance.

Following the official MSRS split, the original test set is kept untouched for final evaluation. We randomly split 10% samples from the original MSRS training set to construct the validation set for hyper-parameter tuning and early-stopping. The remaining 90% samples are used for model training. The identical split is adopted for fusion, semantic segmentation and object-detection tasks to avoid potential data leakage.

A.2 Full Hyper-parameter List

All key hyper-parameters used in TriReward-Fusion are summarized in Table 4. The mean and standard-deviation statistics for reward normalization are collected on the MSRS training set and fixed during inference. No statistics are updated at test time.

Table 4: Summary of hyper-parameters.

Parameter	Value
Optimizer	AdamW
Weight decay	1×10^{-4}
Learning rate (task network)	1×10^{-4}
Learning rate (BLIP backbone)	1×10^{-5}
Warm-up epochs	5
Total epochs (stage-1)	40
Total epochs (stage-2)	80
Batch size	4
ρ (adaptive reward aggregation)	0.1
τ_r (ranking loss temperature)	0.2
m (ranking loss margin)	0.1
α_s (segmentation reward target)	0.7
$\alpha_{box}, \alpha_{cls}, \alpha_{obj}$ (detection reward target)	0.5, 0.3, 0.2
$\lambda_f^r, \lambda_s^r, \lambda_d^r, \lambda_{vl}$ (reward calibration loss weights)	1.0, 1.0, 1.0, 0.2
η (auxiliary supervision weight)	0.4
λ_r (reward guidance loss weight)	0.8
λ_c (cross-task consistency loss weight)	0.6
λ_{pc} (JS divergence weight)	0.2
κ (soft mask sharpness)	20
K_s, K_d (top-K pooling for category presence)	64, 8
τ_v (text-guided attention temperature)	0.07
τ_a (vision-language alignment temperature)	0.1

A.3 Text Prompt Examples for BLIP Category Prototypes

For each shared category $c \in \mathcal{C}_{sd}$, we construct multiple hand-crafted text prompts to compute ensemble text prototypes. Representative prompts are shown below:

- *person*: "a photo of a person", "infrared image containing a person", "visible image with pedestrian"
- *car*: "a photo of a car", "infrared image containing a car", "visible image with vehicle"
- *bicycle*: "a photo of a bicycle", "infrared image containing a bicycle", "visible image with bike"

Multiple prompts for each category are averaged after BLIP text encoder projection to obtain the final category text prototype t_c.

A.4 Degradation Configuration for Fusion Reward Calibration

To generate degraded negative samples $I^{f,-}$ for ranking loss $\mathcal{L}_{rank}^f$, five degradation operations are randomly sampled:

1. Infrared target attenuation: reduce intensity of thermal salient regions.
2. Visible texture suppression: suppress high-frequency texture components.
3. Boundary blurring: apply Gaussian blur on object boundaries.
4. Local over-exposure: increase intensity for local image patches.
5. Contrast degradation: lower global image contrast.

One or more degradation operations are randomly applied with equal probability. Each perturbation degrades at least one desired fusion property while preserving the overall scene content. Degradation magnitude is sampled within predefined value ranges.

B AI USE STATEMENT

In this work, we used generative AI tools for translation assistance. We have not used generative AI tools for generating synthetic datasets, developing theoretical or conceptual frameworks, formulating mathematical claims, assisting proofs, proposing hypotheses, designing research methodology, cleaning datasets, or supporting qualitative data analysis; these required-disclosure tasks are otherwise not applicable to this work. Additionally, we used generative AI tools for drafting parts of the manuscript, improving readability of text, and assisting software code editing. All AI-assisted text and code were fully reviewed, revised and verified by the authors. We take full responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

C ETHICS STATEMENT

All authors of this work have read and strictly adhere to the ICLR Code of Ethics throughout the research and paper submission process. This study does not involve human subjects, confidential personal data, or sensitive private information. All datasets and resources adopted in this work are publicly available and used in full compliance with their official licenses and usage policies. Our proposed method and experimental analyses do not introduce foreseeable harmful biases, discriminatory behaviors, privacy risks, or security threats. No potentially misleading statements, falsified data, or inappropriate claims are presented in this paper. All experimental results and analyses are reported honestly and completely, maintaining full research integrity. Furthermore, there are no conflicts of interest, financial or non-financial sponsorship biases, or ethical concerns associated with this work. We confirm that our research complies with ethical principles including fairness, transparency, accountability, and responsible stewardship of scientific research.

D REPRODUCIBILITY STATEMENT

We have taken measures to facilitate the reproducibility of our experimental results. Implementation details, network configurations, and experimental setups are described in the main paper and appendix. Further details regarding data preprocessing are provided in the supplementary materials. The source code and related experimental data will be made publicly available upon acceptance of this paper. During the anonymous review phase, we do not release the code and dataset to preserve double-blind review. All experimental hyper-parameters and evaluation protocols are thoroughly documented in the manuscript and supplementary materials to support subsequent reproduction.