TFUSE: TEMPORAL FEEDBACK GUIDES CROSS-MODAL SELECTION FOR INFRARED–VISIBLE VIDEO FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared–visible video fusion requires not only high-quality fusion at the individual-frame level, but also temporal consistency across consecutive frames. However, these two objectives may conflict with each other. To address this issue, we propose TFuse. At the objective level, we introduce a learnable loss generator. The generator performs undecimated Haar wavelet decomposition on infrared and visible clips, and predicts frame-adaptive 3×3 kernels for the low- and high-frequency subbands to construct dynamic intensity losses. These losses are learned through online bilevel optimization inspired by the look-ahead principle of Nesterov acceleration. A support clip performs a differentiable virtual update; a query clip evaluates the updated model using a temporal-consistency objective, and the resulting temporal feedback optimizes the loss generator. At the representation level, we introduce Temporal Enhancement with Multi-frequency Propagation Operator (TEMPO). For low-frequency structural information, it jointly models temporal dependencies using clip-level global memory and bidirectional local memory. For high-frequency details, it employs confidence-gated temporal refinement to suppress unreliable temporal propagation. During cross-modal fusion, we introduce Modality-aware Information Routing and Aggregation (MIRA), which combines the basic fused features with the weighted sum of residuals from each modality to enable adaptive information selection. Through these designs, TFuse coordinates the competing objectives, improving temporal coherence while preserving high-quality fused imagery. Code is available in the supplementary material.

1 INTRODUCTION

Infrared (IR) and visible (VI) sensors provide complementary observations of the same scene. Under favorable illumination, visible images capture rich textures, colors, and appearance details, whereas infrared imaging highlights thermal targets and remains informative under low-light conditions. Infrared–visible image fusion therefore aims to integrate complementary information from both modalities into a unified representation, benefiting human perception and downstream vision tasks (Xu et al., 2022; Ma et al., 2022; Zhao et al., 2023a;b; Zhang et al., 2020). In real-world applications, however, these sensors typically acquire synchronized observations as continuous video streams. A desirable video fusion model should therefore produce high-quality, temporally coherent fused frames while exploiting redundancy across neighboring frames. Recent studies have extended infrared–visible fusion to video sequences by modeling cross-frame temporal information (Xie et al., 2024; Guo et al., 2024; Gong et al., 2025; Zhao et al., 2025; Tang et al., 2026).

Despite this progress, a fundamental issue remains in training-objective design, as illustrated in Figure 1. Existing methods typically rely on manually tuned weights for image fusion and temporal consistency losses, although these objectives are not always fully aligned. Image-quality objectives encourage the preservation of salient intensities, textures, gradients, and complementary cross-modal information within each frame, whereas temporal objectives constrain fused outputs across consecutive frames. Since infrared–visible fusion has no unique fusion ground truth, different cross-modal feature-selection strategies may yield similarly high frame-level fusion quality while causing adjacent frames to favor different infrared or visible cues, leading to temporal flicker.

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.

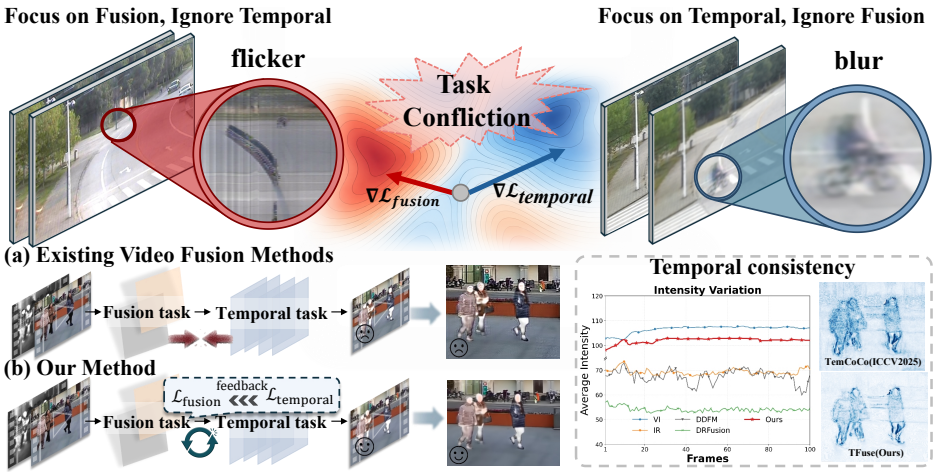


Figure 1: Motivation of TFuse. The existing video fusion methods use inconsistent optimization objectives for fusion loss and temporal loss. We introduce a dynamic fusion loss to alleviate this gradient conflict.

Conversely, strong temporal constraints often rely on predefined assumptions about adjacent-frame evolution, which may become unreliable around motion boundaries, occlusions, or temporally varying cross-modal responses and over-suppress valid details. More importantly, fixed loss weights only change optimization strengths and cannot use temporal evidence to modify how the fusion objective selects cross-modal information. This motivates adapting the frame-level image-quality objective using the temporal behavior of the fused video.

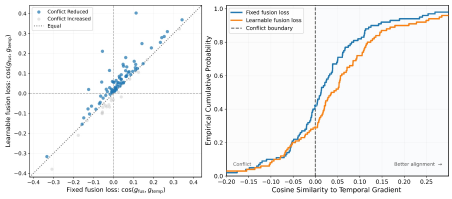


Figure 2: Left: Fusion-temporal gradient conflict; Right: Fusion-temporal gradient cosine distribution.

To address this issue, we propose TFuse, which exploits temporal information at both the objective and representation levels to guide cross-modal information selection and cross-frame propagation. At the objective level, we introduce a frequency-aware learnable fusion-loss generator and use online bilevel optimization to let subsequent temporal behavior guide the frame-level fusion objective. In this way, the image-quality criterion can adapt according to its impact on temporal consistency, thereby alleviating the conflict between fusion quality and temporal consistency. At the representation level, we further introduce MIRA and TEMPO to model the integrate of features and residuals of each modality and the temporal dependencies of different frequency components, respectively. Through these designs, TFuse coordinates current-frame information selection and cross-frame propagation, improving temporal consistency while preserving high-quality fusion results.

Our main contributions are summarized as follows:

- We propose an online bilevel optimization framework for infrared–visible video fusion to coordinate single-frame fusion quality and temporal consistency through a learnable dynamic fusion objective. The framework uses subsequent temporal behavior to guide the frame-level fusion objective, allowing the fusion criterion to adapt to its impact on temporal consistency and thereby alleviating the conflict between the two objectives.
- We propose TEMPO, which performs selective temporal modeling according to the characteristics of different frequency components. The module combines global–local dual memory to model low-frequency temporal dependencies and employs confidence-gated temporal refinement to suppress unreliable propagation of high-frequency details, thereby improving temporal stability.
- We design MIRA that explicitly models integrates features and residuals of each modality to enable adaptive cross-modal information selection, thereby improving the quality and reliability of fused representations.

2 Related Work

Infrared-visible image and video fusion. Deep image fusion has evolved from representation learning (Xu et al., 2022; Li et al., 2023) to decomposition and diffusion models (Zhao et al., 2023a;b). MURF couples registration with fusion (Xu et al., 2023), while recent general fusion models explore task-customized adapters (Zhu et al., 2024), task-guided architecture learning (Liu et al., 2024), and low-level task interaction (Cheng et al., 2025). Video fusion further exploits temporal dependencies. RCVS jointly addresses registration and fusion (Xie et al., 2024), and DFTP models temporal phase cues (Guo et al., 2024). TemCoCo introduces explicit temporal modeling (Gong et al., 2025), UniVF adopts multi-frame learning (Zhao et al., 2025), and VideoFusion models cross-modal and temporal contexts (Tang et al., 2026). Recent approaches explore streaming diffusion (Liu et al., 2026), drift-resilient temporal modeling (Li et al., 2026b), uncertainty-guided spatiotemporal consistency (Zhao et al., 2026), and motion-aware sparse interaction (Li et al., 2026a).

Temporal modeling with frequency decomposition. Temporal aggregation should exploit cross-frame redundancy without oversmoothing or unreliable propagation. BasicVSR and BasicVSR++ employ bidirectional propagation and feature alignment (Chan et al., 2021; 2022), while VRT and RVRT model longer-range dependencies through Transformer interactions (Liang et al., 2024; 2022). Frequency-domain studies further reveal temporal priors: Zhu et al. (2023) explore frequency-temporal characteristics for video deblurring, while VD-Diff distinguishes low-frequency structure from high-frequency details through wavelet-aware restoration (Rao et al., 2024). In fusion, low-frequency components mainly represent illumination and target structure, whereas high-frequency bands encode motion-sensitive edges and textures. TEMPO therefore uses broader temporal context for low-frequency components and reliability-aware propagation for high-frequency details.

Bilevel optimization. Bilevel optimization formulates nested lower- and upper-level problems, where upper-level variables depend on lower-level optimization dynamics. Gradient-based approaches differentiate through iterative learning to obtain hypergradients (Maclaurin et al., 2015; Franceschi et al., 2018), while approximate-gradient methods reduce the cost of solving the inner problem (Pedregosa, 2016). Truncated back-propagation further approximates hypergradients through limited lower-level steps (Shaban et al., 2019; Kwon et al., 2024). Nesterov acceleration introduces a look-ahead principle that evaluates gradients at anticipated iterates, incorporating prospective optimization dynamics into parameter updates (Nesterov, 1983; 2004).

3 Method

3.1 Overview

Infrared–visible video fusion aims to integrate complementary information from two temporally aligned video sequences into a fused sequence while maintaining consistency across frames. Given an infrared video $\mathbf{I} = \{\mathbf{I}_t\}_{t=1}^{T}$ and a visible video $\mathbf{V} = \{\mathbf{V}_t\}_{t=1}^{T}$, where t indexes the frame and T denotes the sequence length, our goal is to generate a fused video $\mathbf{F} = \{\mathbf{F}_t\}_{t=1}^{T}$. Each fused frame $\mathbf{F}_t$ should preserve complementary information from the corresponding infrared and visible frames.

We propose TFuse, a unified framework that coordinates cross-modal fusion and temporal consistency. During training, a loss generator g_ϕ produces frame-adaptive fusion losses and is optimized using feedback from the temporal-consistency objective, alleviating the optimization conflict between fusion quality and temporal consistency. At the representation level, cross-modal fusion and temporal modeling are carried out across three resolution levels, where MIRA adaptively selects information from the IR and VI modalities, while TEMPO enables frequency-aware propagation across frames. Together, these components learn what cross-modal information to fuse and which temporal context to exploit for temporally consistent fusion. Figure 3 illustrates the overall framework.

3.2 Temporally Guided Learnable Intensity Loss

Dynamic Fusion Loss Construction. Given infrared and visible video clips, we construct a learnable dynamic fusion loss by separately modeling low- and high-frequency components, parameterized by frame-wise 3×3 dynamic loss kernels. Haar SWT decomposes the single-channel representations of the infrared and visible clips into $\mathbf{L}^{(m)}$ and $\mathbf{H}^{(m,h)}$, where $m \in \{\mathrm{IR}, \mathrm{VI}\}$ and $h \in \{\mathrm{lh}, \mathrm{hl}, \mathrm{hh}\}$.

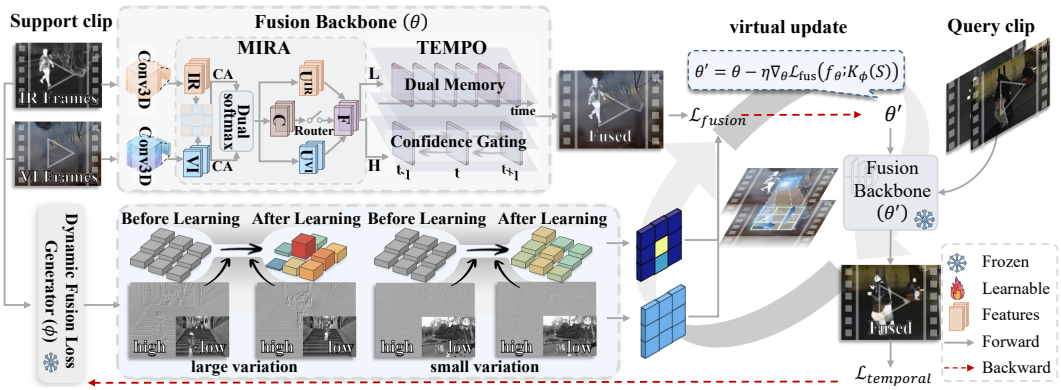


Figure 3: Overview of TFuse

Two independent 3D encoders process the corresponding IR and VI wavelet components and their cross-modal differences to predict frame-wise 3×3 kernel adjustments. The low- and high-frequency kernels are constructed from Gaussian and Laplacian priors $\mathbf{K}_G$ and $\mathbf{K}_\Delta$, respectively:

$$K^L_{t,ij} = \frac{K_{G,ij} \exp\left(P^L_{t,ij}\right)}{\sum_{p,q} K_{G,pq} \exp\left(P^L_{t,pq}\right)}. \tag{1}$$

$$\mathbf{K}^H_t = \frac{\mathbf{K}_\Delta + \mathbf{P}^H_t - \text{mean}\left(\mathbf{K}_\Delta + \mathbf{P}^H_t\right)}{\left\|\mathbf{K}_\Delta + \mathbf{P}^H_t - \text{mean}\left(\mathbf{K}_\Delta + \mathbf{P}^H_t\right)\right\|_1}. \tag{2}$$

Here, $\mathbf{P}^{L/H}_t$ denotes the kernel adjustments, while $\mathbf{K}^{L/H}_t$ denotes the corresponding dynamic kernels, collectively denoted by $\mathbf{K}^{L/H}$; $\exp(\cdot)$ is the exponential function, and i, j, p, q index kernel positions.

Fusion Objective Formulation. The low-frequency dynamic intensity loss is defined as

$$\mathcal{L}^L_{\text{int}} = \left\|\mathbf{K}^L * \mathbf{L}^{(F)} - \max\left(\mathbf{K}^L * \mathbf{L}^{(\text{IR})}, \mathbf{K}^L * \mathbf{L}^{(\text{VI})}\right)\right\|_1. \tag{3}$$

Here, $*$ denotes frame-wise dynamic convolution. The low-frequency kernels control local source-information selection, while temporal feedback adjusts their weights through bilevel optimization.

The corresponding high-frequency loss is

$$\mathcal{L}^H_{\text{int}} = \frac{1}{3} \sum_{h \in \{\text{lh,hl,hh}\}} \left\|\mathbf{K}^H * \mathbf{H}^{(F,h)} - \text{MaxAbs}\left(\mathbf{K}^H * \mathbf{H}^{(\text{IR},h)}, \mathbf{K}^H * \mathbf{H}^{(\text{VI},h)}\right)\right\|_1. \tag{4}$$

$\text{MaxAbs}(u, v)$ selects the response with the larger magnitude while preserving its sign. The high-frequency kernels adapt to video content, determining which local high-frequency structures receive stronger responses under the current temporal context.

3.3 Temporal Feedback Guided Bilevel Optimization

To better learn dynamic fusion objectives, we optimize the loss generator g_ϕ within an online bilevel optimization framework. At each iteration, the dynamic fusion loss first performs a differentiable virtual update of the fusion network and is then optimized using temporal feedback from the updated fused frames.

Virtual Fusion Update. Given a support clip $\mathbf{S}$ and a query clip $\mathbf{Q}$, let $\theta_s \subset \theta$ denote the parameters involved in the virtual update, including the three-scale MIRA modules and the fusion decoder. The loss generator produces low- and high-frequency kernels $\mathbf{K}_\phi(\mathbf{S}) = \{\mathbf{K}^L_\phi(\mathbf{S}), \mathbf{K}^H_\phi(\mathbf{S})\}$, and the support loss performs a one-step virtual update:

$$\theta'_s = \theta_s - \eta \nabla_{\theta_s} \mathcal{L}_{\text{fus}}\left(f_\theta(\mathbf{S}); \mathbf{K}_\phi(\mathbf{S})\right), \tag{5}$$

where η is the virtual step size.

Temporal Guidance and Gradient Propagation. Let $\theta'(\phi)$ denote the virtual network parameters obtained by replacing θ_s with $\theta'_s(\phi)$. The query clip evaluates this update through $\mathcal{L}_{\text{temp}}\left(f_{\theta'(\phi)}(\mathbf{Q}), \mathbf{Q}\right)$. We establish a differentiable gradient pathway from the temporal loss to the dynamic fusion loss through the virtual update. The dynamic fusion loss on the support clip is denoted as

$$\mathcal{L}_{\text{fus}}(\theta_s, \phi) \triangleq \mathcal{L}_{\text{fus}}\left(f_\theta(\mathbf{S}); \mathbf{K}_\phi(\mathbf{S})\right). \tag{6}$$

Since the temporal loss on the query clip depends on the loss generator parameters ϕ through the virtual parameters $\theta'_s(\phi)$, the chain rule gives

$$\nabla_\phi \mathcal{L}_{\text{temp}} = \left(\frac{\partial \theta'_s}{\partial \phi}\right)^\top \nabla_{\theta'_s} \mathcal{L}_{\text{temp}}. \tag{7}$$

Differentiating the virtual update in Eq. (5) with respect to ϕ, we obtain

$$\frac{\partial \theta'_s}{\partial \phi} = \frac{\partial}{\partial \phi}\left[\theta_s - \eta \nabla_{\theta_s} \mathcal{L}_{\text{fus}}(\theta_s, \phi)\right]. \tag{8}$$

During the current virtual update, the fusion parameters θ_s before the update are independent of ϕ, and therefore $\frac{\partial \theta_s}{\partial \phi} = 0$. Using this independence property in Eq. (8) gives

$$\frac{\partial \theta'_s}{\partial \phi} = -\eta \frac{\partial}{\partial \phi}\left(\nabla_{\theta_s} \mathcal{L}_{\text{fus}}\right), \tag{9}$$

Equivalently, Eq. (9) can be written as

$$\frac{\partial \theta'_s}{\partial \phi} = -\eta \frac{\partial^2 \mathcal{L}_{\text{fus}}}{\partial \phi \, \partial \theta_s}. \tag{10}$$

Substituting Eq. (10) into Eq. (7) yields the gradient for updating the loss generator parameters ϕ:

$$\nabla_\phi \mathcal{L}_{\text{temp}} = -\eta \left(\frac{\partial^2 \mathcal{L}_{\text{fus}}}{\partial \phi \, \partial \theta_s}\right)^\top \nabla_{\theta'_s} \mathcal{L}_{\text{temp}}. \tag{11}$$

Unlike directly optimizing a fixed weighted sum of $\mathcal{L}_{\text{fus}}$ and $\mathcal{L}_{\text{temp}}$, temporal feedback here reshapes the fusion objective itself. By differentiating through the virtual update, $\mathcal{L}_{\text{temp}}$ updates the loss-generator parameters and thereby adapts the subsequent low- and high-frequency fusion criteria. The loss generator is used only during training and introduces no additional inference-time parameters. The second-order term in Eq. (10) denotes the mixed derivative of the dynamic fusion loss with respect to the fusion network parameters θ_s and the loss generator parameters ϕ. This gradient allows the temporal loss on the query clip to propagate back to the dynamic fusion loss generator through the differentiable virtual update, enabling the fusion objective to be adaptively adjusted according to the actual temporal performance of the virtually updated network. After completing the above update, g_ϕ is frozen and used to guide the standard training of the fusion network parameters. The detailed procedure of the bilevel optimization is provided in Algorithm 1 of the supplementary material.

3.4 Temporal Enhancement with Multi-frequency Propagation Operator

The temporal information exhibits different characteristics in different frequency components. Low-frequency components mainly encode scene structure and illumination, whereas high-frequency components capture edges and textures and are more sensitive to motion and misalignment. Based on this observation, we propose Temporal Enhancement with Multi-frequency Propagation Operator (TEMPO), which performs temporal information propagation in a frequency-aware manner. Haar SWT decomposes $\mathbf{X} \in \mathbb{R}^{B\times T\times C\times H\times W}$ into $\mathbf{L}$, $\mathbf{H}^{lh}$, $\mathbf{H}^{hl}$, $\mathbf{H}^{hh}$, where B, T, C, H, W denote batch, time, channel, height, and width.

Low-frequency memory. For low-frequency features, TEMPO aggregates complementary temporal information through a dual-memory mechanism. Specifically, the global memory module captures long-range dependencies within the entire video segment through an attention-based aggregation approach, while the local memory module models the short-range bidirectional correlations between adjacent frames. Subsequently, a lightweight routing function predicts adaptive weights based on the



Figure 4: Motion-Compensated Temporal Residuals on HDO for Temporal Consistency Evaluation.

current features and the reliability of different memory information, thereby obtaining the enhanced low-frequency representation:

$$\mathbf{M}_t = \rho_t^G \mathbf{M}_t^G + \rho_t^L \mathbf{M}_t^L. \tag{12}$$

Here, $\mathbf{M}_t^{G/L}$ denote the global/local memories and $\rho_t^{G/L}$ their routing weights.

Frequency-selective update. For high-frequency components, long-range propagation may introduce artifacts due to inaccurate inter-frame correspondences. Therefore, TEMPO restricts high-frequency propagation to adjacent frames:

$$\mathbf{X}_t' = \mathbf{X}_t + \alpha_t^L \left(\widehat{\mathbf{L}}_t - \mathbf{L}_t \right) + \sum_{h \in \{lh, hl, hh\}} \alpha_t^h \left(\widehat{\mathbf{H}}_t^h - \mathbf{H}_t^h \right). \tag{13}$$

Among them, α_t^L and α_t^h are adaptive gates, while $\widehat{\mathbf{L}}_t$ and $\widehat{\mathbf{H}}_t^h$ denote temporally propagated features. This frequency-specific design captures long-range temporal dependencies in low-frequency structures while preserving rapidly varying high-frequency details.This frequency-aware strategy improves temporal consistency while preserving spatial details.

3.5 Modality-aware Information Routing and Aggregation

To obtain more effective cross-modal representations, we introduce MIRA, which performs reliability-aware fusion and residual routing within local windows. Within each window, the spatially averaged IR and VI value features are concatenated to summarize the local cross-modal context. To extract comparable information from both modalities, we introduce a set of learnable shared queries as a common retrieval basis, which are adapted using the local context. The adapted queries independently retrieve IR and VI observations through cross-attention. Their preliminary observations and absolute difference further refine the shared queries for a second retrieval, yielding O^{IR} and O^{VI}. Bidirectional softmax and learnable null tokens suppress unsupported matches, while reliability-aware modality weights combine the retrieved observations to construct the fusion reference C:

$$\mathbf{C} = \omega^{\mathrm{IR}} \odot \mathbf{O}^{\mathrm{IR}} + \omega^{\mathrm{VI}} \odot \mathbf{O}^{\mathrm{VI}}. \tag{14}$$

Here, $\mathbf{C}$ is the shared representation, $\omega^{\mathrm{IR}}, \omega^{\mathrm{VI}}$ are the IR/VI adaptive weights, and $\odot$ denotes element-wise multiplication.To further refine the fusion, we extract modality-wise residuals $\mathbf{U}^{\mathrm{IR}}$ and $\mathbf{U}^{\mathrm{VI}}$ relative to $\mathbf{C}$, and adaptively route them for final fusion:

$$\mathbf{U}^m = \psi_m \left(\mathbf{O}^m - \mathbf{C} \right), \qquad \mathbf{X}^F = r_c \odot \mathbf{C} + r_{\mathrm{IR}} \odot \mathbf{U}^{\mathrm{IR}} + r_{\mathrm{VI}} \odot \mathbf{U}^{\mathrm{VI}}, \qquad m \in \{\mathrm{IR}, \mathrm{VI}\}. \tag{15}$$

Here, ψ_m denotes a modality-specific transformation, while r_c, r_{IR}, and r_{VI} adaptively weight the fusion representation and modality residuals.This reference-residual decomposition preserves complementary modality-specific information that is not fully captured by the shared fusion reference.

3.6 Loss Function

We combine the learnable intensity losses with standard fusion constraints to form the spatial fusion objective:

$$\mathcal{L}_{\mathrm{fus}} = \lambda_L \mathcal{L}_{\mathrm{int}}^L + \lambda_H \mathcal{L}_{\mathrm{int}}^H + \lambda_g \mathcal{L}_{\mathrm{grad}} + \lambda_c \mathcal{L}_{\mathrm{color}} + \lambda_s \mathcal{L}_{\mathrm{SSIM}}. \tag{16}$$

Table 1: Quantitative comparison on M3SVD and HDO. The best and second-best results are indicated by boldface and underlining, respectively.

Methods	Multi-Modal Multi-Scene Video Dataset (M3SVD)						High-quality Dual-Optical Dataset (HDO)					
	EN↑	MI↑	SSIM↑	VIF↑	flowD↓	BiSWE↓	EN↑	MI↑	SSIM↑	VIF↑	flowD↓	BiSWE↓
U2Fusion (Xu et al., 2022)	6.064	1.771	0.519	0.278	6.295	5.907	6.609	2.333	0.626	0.252	7.245	5.708
LRRNet (Li et al., 2023)	6.809	2.892	0.552	0.374	6.451	7.174	6.661	2.542	0.598	0.230	6.202	6.672
DDFM (Zhao et al., 2023b)	6.717	2.655	0.580	0.426	5.551	6.050	6.849	2.813	0.645	0.306	5.960	5.772
TIMFusion (Liu et al., 2024)	6.696	2.511	0.488	0.334	5.875	6.760	6.327	2.438	0.502	0.225	5.922	7.865
TC-MOA (Zhu et al., 2024)	7.236	2.886	0.564	0.466	5.252	6.630	7.138	2.912	**0.653**	0.316	7.392	5.824
GIFNet (Cheng et al., 2025)	6.782	2.364	0.565	0.354	6.355	6.894	6.963	2.269	0.597	0.235	9.601	8.560
TemCoCo (Gong et al., 2025)	7.266	2.897	0.522	0.432	4.538	5.410	7.115	2.707	0.592	0.253	7.345	5.185
DRFusion (Li et al., 2026b)	6.458	2.294	0.563	0.370	4.471	5.232	6.899	2.958	0.631	0.305	8.556	7.516
SDMFusion (Liu et al., 2026)	7.253	2.589	0.512	0.466	4.173	4.049	**7.287**	2.617	0.623	0.349	8.090	7.416
MAVFusion (Li et al., 2026a)	7.249	2.894	0.548	0.488	4.065	4.683	7.213	3.327	0.636	0.414	7.863	5.833
VideoFusion (Tang et al., 2026)	7.113	3.691	0.606	0.524	3.559	2.445	7.040	3.153	0.628	0.538	5.907	3.911
Ours	**7.319**	**3.980**	**0.627**	**0.758**	**2.373**	**1.206**	7.227	**3.338**	0.635	**0.540**	**4.691**	**2.916**

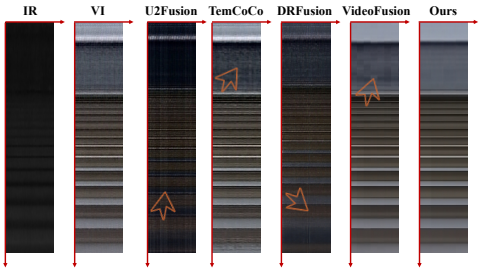


Figure 5: Temporal consistency of source and fused videos following (Pan et al., 2023).

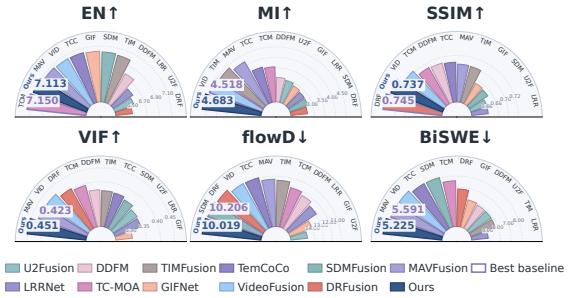


Figure 6: Quantitative comparison on VTMOT.

$\mathcal{L}_{\text{int}}^{L}$ and $\mathcal{L}_{\text{int}}^{H}$ are learnable losses, $\mathcal{L}_{\text{grad}}$ (Zhao et al., 2024) and $\mathcal{L}_{\text{color}}$ (Wang et al., 2024) follow standard fusion constraints, while $\mathcal{L}_{\text{SSIM}}$ (Wang et al., 2004) adopts structural similarity. The temporal objective is

$$\mathcal{L}_{\text{temp}} = \sum_{\mathbf{Z}\in\{\mathbf{I},\mathbf{V}\}} \lambda_{\mathbf{Z}} \left[\|\Delta\mathbf{F} - \Delta\mathbf{Z}\|_1 + \lambda_{\cos} \left(1 - \cos\left(\Delta\mathbf{F}, \Delta\mathbf{Z}\right)\right)\right]. \tag{17}$$

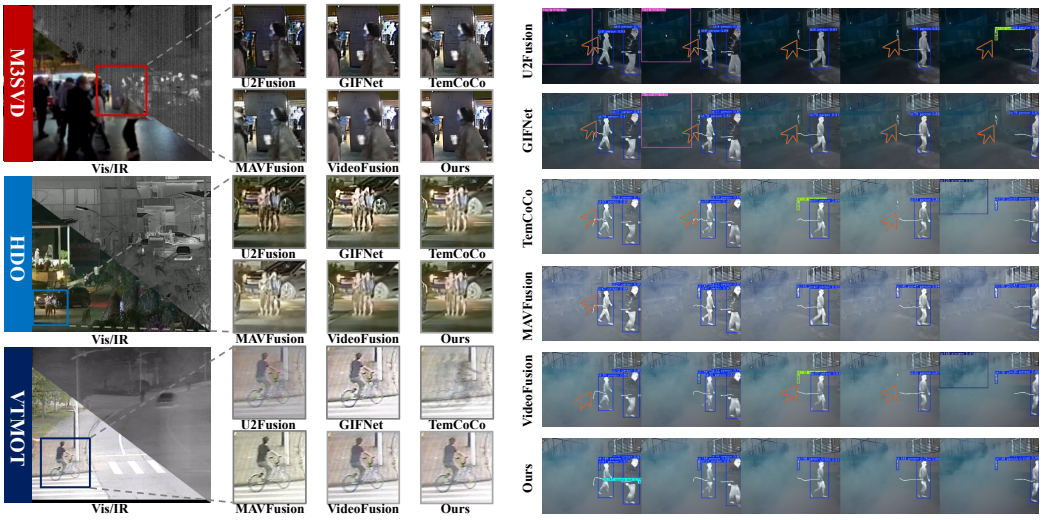
Here, Δ denotes the inter-frame difference and $\lambda_{\cos}$ weights the cosine term.

4 Experiments

4.1 Experimental Setup

We evaluate TFuse on M3SVD, HDO, and VTMOT. For M3SVD and HDO, we follow the experimental setup of VideoFusion (Tang et al., 2026). For VTMOT, we randomly select 30 videos, using 5 for testing and the remaining 25 for training. All methods are retrained and evaluated using the same data split and experimental settings. We report EN (Roberts et al., 2008), MI (Qu et al., 2002), SSIM (Wang et al., 2004), and VIF (Han et al., 2013) for fusion quality, and flowD (Gong et al., 2025) and BiSWE (Zhao et al., 2025) for temporal consistency.We compare TFuse with eleven SOTA methods: U2Fusion (Xu et al., 2022), LRRNet (Li et al., 2023), DDFM (Zhao et al., 2023b), TC-MoA (Zhu et al., 2024), TIMFusion (Liu et al., 2024), GIFNet (Cheng et al., 2025), TemCoCo (Gong et al., 2025), VideoFusion (Tang et al., 2026), SDMFusion (Liu et al., 2026), DRFusion (Li et al., 2026b), and MAVFusion (Li et al., 2026a).

We set T = 7 during training and T = 25 for testing. The three scales use channels $[32, 64, 128]$ and $[2, 2, 4]$ backbone blocks; TEMPO uses $[1, 1, 2]$ temporal blocks.The hyperparameters balancing various losses are empirically set as $\lambda_L = 15$, $\lambda_H = 1$, $\lambda_g = 1$, $\lambda_c = 100$, $\lambda_s = 1$, $\lambda_I = 0.1$, $\lambda_V = 0.4$, and $\lambda_{\cos} = 0.2$. The fusion network and loss generator are optimized using AdamW with



(a) Fusion comparison visualization (b) Tracking visualization

Figure 7: Qualitative comparison of fusion and downstream tracking.

$\beta_1 = 0.9$ and $\beta_2 = 0.999$, with initial learning rates of 1×10^{-4} and 5×10^{-5}, respectively, for 20 epochs under a cosine annealing schedule. The bilevel update is performed every four iterations with inner learning rate 10^{-4}.Our method is implemented in PyTorch. All experiments are conducted on four NVIDIA RTX PRO 6000 GPUs and a 2.10 GHz Intel Xeon Gold 6530 CPU.

4.2 Video Fusion Experiments

Fusion Performance Comparison. Quantitatively, Table 1 and Figure 6 show that TFuse achieves SOTA fusion quality across all three datasets, ranking first on all four M3SVD metrics and achieving the best MI and VIF on HDO and VTMOT. Qualitatively, Figure 7(a) shows that competing methods suffer from blurred details on M3SVD and HDO or color distortion on VTMOT, while TFuse better preserves texture structures and appearance. Overall, TFuse effectively integrates complementary infrared and visible information, producing clearer and more informative fusion results.

Temporal Consistency Comparison. The quantitative results are shown in Table 1 and Figure 6. TFuse achieved the lowest flowD and BiSWE on all three datasets, indicating its superior temporal stability and consistency. In terms of qualitative results, Figure 4 visualizes the motion compensation time residuals of four consecutive frames. We aligned each previous frame with the current frame using optical flow to reduce the impact of motion, and then calculated and averaged the residuals between adjacent frames. A lower residual response indicates better cross-frame consistency.

Compared with other methods, the residual response of TFuse is weaker, especially in static background areas, demonstrating better temporal consistency. Figure 5 further shows that other methods have obvious flickering and temporal discontinuity, while TFuse can generate a smoother and continuous temporal structure; Figure 8 also verifies this with more stable frame-by-frame average brightness changes. Thanks to TEMPO and the double-layer loss learning, TFuse can effectively utilize time-dependent relationships and time feedback to suppress cross-frame fluctuations. Figure 7(a) shows fewer artifacts, further indicating that the improvement in temporal consistency does not come at the expense of spatial fusion quality.

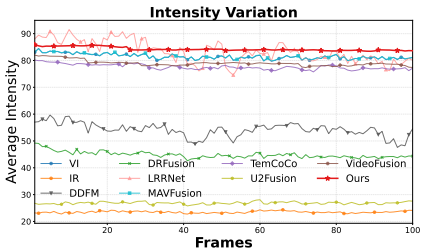


Figure 8: Temporal variation of column-averaged intensity at a fixed spatial position.

Table 2: Progressive ablation of TEMPO, MIRA, and bilevel loss learning from the VideoFusion baseline on M3SVD and HDO.

Configuration	M3SVD						HDO					
	EN↑	VIF↑	MI↑	SSIM↑	flowD↓	BiSWE↓	EN↑	VIF↑	MI↑	SSIM↑	flowD↓	BiSWE↓
VideoFusion	7.113	0.524	3.691	0.606	3.559	2.445	7.040	0.538	3.153	0.628	5.907	3.911
+ TEMPO	7.165	0.529	3.710	0.611	3.098	1.925	7.060	0.535	3.192	0.629	5.425	3.585
+ TEMPO + MIRA	7.295	0.642	3.915	0.621	3.006	1.831	7.217	0.539	3.308	0.632	5.214	3.431
+ TEMPO + Bilevel loss learning	7.175	0.602	3.810	0.614	2.798	1.525	7.128	0.538	3.239	0.630	4.956	3.194
Full model (TFuse)	**7.319**	**0.758**	**3.980**	**0.627**	**2.373**	**1.206**	**7.227**	**0.540**	**3.338**	**0.635**	**4.691**	**2.916**

4.3 Ablation Studies

To identify the sources of performance gains over the baseline and investigate the individual contributions and complementarity of TEMPO, MIRA, and bilevel loss learning, we conduct progressive ablation studies on M3SVD and HDO. As shown in Table 2, we start from VideoFusion as a unified baseline, first introduce TEMPO, then incorporate either MIRA or bilevel loss learning on top of TEMPO, and finally combine all three components to obtain the full TFuse. Introducing TEMPO consistently reduces flowD and BiSWE on both datasets, demonstrating that adaptive temporal modeling of low- and high-frequency information effectively improves cross-frame consistency. Adding MIRA further improves fusion quality, increasing VIF/MI from 0.529/3.710 to 0.642/3.915 on M3SVD, which validates its effectiveness in cross-modal feature fusion. Bilevel loss learning, in turn, brings more pronounced temporal gains, further reducing flowD/BiSWE to 2.798/1.525 on M3SVD, with a similar trend observed on HDO, demonstrating the effectiveness of using temporal feedback to guide the optimization of the fusion objective. Finally, the full model achieves the best results across all metrics on both datasets, showing that TEMPO, MIRA, and bilevel loss learning play complementary roles in low- and high-frequency cross-frame information modeling, cross-modal feature fusion, and temporal-feedback-guided optimization, respectively.

4.4 Downstream Tracking

Beyond perceptual quality, effective video fusion should also provide reliable representations for downstream video perception. To further evaluate the utility of fused videos for downstream vision tasks, we conduct object tracking experiments, as shown in Fig. 7(b). All fused videos are fed into the same pretrained tracker under identical inference settings, without task-specific fine-tuning. We conduct object tracking experiments on the VTMOT dataset, using a pre-trained YOLO v11 (Khanam & Hussain, 2024) as the baseline tracker. Compared with competing methods that suffer from missed detections, false responses, or unstable tracking across consecutive frames, TFuse maintains more continuous target responses and stable tracking trajectories. These results demonstrate that TFuse not only improves fusion quality and temporal consistency, but also provides more reliable representations for downstream video perception.

5 Conclusion

We presented TFuse, an infrared-visible video fusion framework that coordinates cross-modal fusion and temporal consistency through temporal-feedback-guided bilevel learning. Rather than treating fusion quality and temporal consistency as independently weighted objectives, TFuse leverages the temporal behavior of fused videos to guide the learning of the fusion objective itself. Specifically, we introduced a frequency-aware learnable intensity loss adaptively shaped by temporal feedback from the fused video. Meanwhile, MIRA progressively routes and aggregates cross-modal information for adaptive fusion, while TEMPO explicitly models low- and high-frequency components with tailored cross-frame propagation. Extensive experiments on multiple infrared-visible video datasets demonstrate that TFuse improves temporal coherence while preserving high-quality fusion results. Ablation studies further validate the complementary effects of bilevel loss learning, MIRA, and TEMPO, while downstream tracking experiments demonstrate the effectiveness of the fused videos for subsequent video perception. These results demonstrate that temporal information can serve not only as a constraint on inter-frame stability, but also as effective feedback for learning the fusion objective itself.

AI USE STATEMENT

This research utilized AI assistance for code debugging and grammatical refinement. All experimental designs, analyses, results, and conclusions were developed independently by the authors without generative AI input. We employed AI tools solely for technical implementation support and language polishing to ensure clear communication of our findings.

REFERENCES

Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4947–4956, 2021.

Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5972–5981, 2022.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for ALL: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28102–28112, 2025.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1568–1577, 2018.

Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14326–14335, 2025.

Xiaoming Guo, Fengbao Yang, and Linna Ji. Dftp: A mimic fusion method driven by difference feature time-phase perception for infrared and visible video. *IEEE Transactions on Instrumentation and Measurement*, 73:1–20, 2024.

Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. doi: 10.1016/j.inffus.2011.08.002.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, volume 2024, pp. 47473–47518, 2024.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023. doi: 10.1109/TPAMI.2023.3268209.

Xilai Li, Weijun Jiang, Xiaosong Li, Yang Liu, Hongbin Wang, Tao Ye, Huafeng Li, and Haishu Tan. Mavfusion: Efficient infrared and visible video fusion via motion-aware sparse interaction. In *Computer Vision – ECCV 2026*, pp. 641–658. Springer Nature Switzerland, 2026a.

Xingyuan Li, Haoyuan Xu, Shulin Li, Xiang Chen, Zhiying Jiang, and Jinyuan Liu. Drfusion: Drift-resilient temporally consistent infrared–visible video fusion. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024. doi: 10.1109/TIP.2024.3372454.

Jinyuan Liu, Ludan Sun, Tengyu Ma, Chunyan Yang, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Streaming diffusion model for fast infrared and visible video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14305–14314, 2026.

Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2113–2122, 2015.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2004.

Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22191–22200. IEEE, 2023.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 737–746, 2016.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313–315, 2002. doi: 10.1049/el:20020212.

Chen Rao, Guangyuan Li, Zehua Lan, Jiakai Sun, Junsheng Luan, Wei Xing, Lei Zhao, Huaizhong Lin, Jianfeng Dong, and Dalong Zhang. Rethinking video deblurring with wavelet-aware dynamic transformer and diffusion model. In *Proceedings of the European Conference on Computer Vision*, pp. 421–437, 2024.

J. Wesley Roberts, Jan A. van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. doi: 10.1117/1.2945910.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 1723–1732, 2019.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: a spatio-temporal collaborative network for multi-modal video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19559–19569, 2026.

Hebaixu Wang, Hao Zhang, Xunpeng Yi, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. Terf: Text-driven and region-aware flexible visible and infrared image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 935–944, 2024.

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.

Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. Rcvs: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*, 26:11031–11043, 2024.

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, 2022.

Han Xu, Jiteng Yuan, and Jiayi Ma. MURF: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. doi: 10.1109/TPAMI.2023.3283682.

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*, pp. 12797–12804, 2020.

Cheng Zhao, Tianyun Song, Zhiliang Wu, Tianfu Wang, Moncef Gabbouj, Guanghui Yue, Baiying Lei, and Wei Zhou. Uncertainty-guided spatiotemporal consistency fusion network for infrared-visible video fusion under extremely low-light conditions. *IEEE Transactions on Image Processing*, 35:8894–8909, 2026. doi: 10.1109/TIP.2026.3719477.

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, 2023a.

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

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 41st International Conference on Machine Learning*, pp. 60749–60765, 2024.

Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. In *Advances in Neural Information Processing Systems*, 2025.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7099–7108, 2024.

Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, and Feng Zhao. Exploring temporal frequency spectrum in deep video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12428–12437, 2023.

A Supplementary Method Details

A.1 Detailed Bilevel Optimization Procedure

Unlike static image fusion, consecutive video frames provide temporal evolution information that reveals whether the current fusion objective leads to temporally stable cross-modal information selection. We exploit this video-specific supervision to adapt the frame-level fusion objective according to its subsequent temporal behavior.

Inspired by the look-ahead principle of Nesterov acceleration, we evaluate the consequence of the current fusion objective through a differentiable virtual update before applying it to the actual network update. Specifically, a support clip $\mathbf{S}$ is first used to construct the dynamic fusion objective and perform a one-step differentiable look-ahead update of the selected fusion parameters. The virtually updated network is then evaluated on a temporally ordered query clip $\mathbf{Q}$, whose temporal behavior serves as a look-ahead signal for assessing the effect of the current fusion objective.

By backpropagating this temporal feedback through the virtual update, the dynamic loss generator learns to adjust the frame-level fusion objective according to its impact on subsequent temporal behavior. In this way, temporal evolution unique to video is not merely imposed as an additional regularization term, but is used as feedback to shape how infrared and visible information is selected during fusion. After the bilevel update, the loss generator is frozen and used to guide the standard training of the fusion network. Algorithm 1 summarizes the complete procedure.

Algorithm 1 Temporal-Feedback-Guided Bilevel Optimization

Require: Fusion network f_θ; dynamic loss generator g_ϕ; virtual step size η; network learning rate η_θ; loss-generator learning rate γ; bilevel-update interval M
Ensure: Trained fusion network f_θ

1: **for** each training iteration n **do**
2: **if** $n \bmod M = 0$ **then**
3: Sample a support clip $\mathbf{S}$ and a temporally ordered query clip $\mathbf{Q}$.
4: Generate frequency-dependent dynamic kernels:

$$\mathbf{K}_\phi(\mathbf{S}) = \left\{\mathbf{K}_\phi^L(\mathbf{S}), \mathbf{K}_\phi^H(\mathbf{S})\right\}.$$

5: Compute the fused support clip:

$$\mathbf{F}_\mathbf{S} = f_\theta(\mathbf{S}).$$

6: Construct the dynamic fusion objective:

$$\mathcal{L}_{\text{fus}}^{\mathbf{S}} = \mathcal{L}_{\text{fus}}\left(\mathbf{F}_\mathbf{S}; \mathbf{K}_\phi(\mathbf{S})\right).$$

7: Select the fusion parameters $\theta_s \subset \theta$ involved in the virtual update.
8: Perform a one-step differentiable look-ahead update:

$$\theta_s' = \theta_s - \eta \nabla_{\theta_s} \mathcal{L}_{\text{fus}}^{\mathbf{S}}.$$

9: Construct the virtual parameters $\theta'(\phi)$ by replacing θ_s with θ_s' while keeping the remaining parameters unchanged.
10: Evaluate the look-ahead temporal behavior:

$$\mathbf{F}_\mathbf{Q}' = f_{\theta'(\phi)}(\mathbf{Q}).$$

11: Compute the temporal feedback objective:

$$\mathcal{L}_{\text{temp}}^{\mathbf{Q}} = \mathcal{L}_{\text{temp}}\left(\mathbf{F}_\mathbf{Q}', \mathbf{Q}\right).$$

12: Backpropagate the temporal feedback through the differentiable look-ahead update:

$$\nabla_\phi \mathcal{L}_{\text{temp}}^{\mathbf{Q}} = -\eta \left(\frac{\partial^2 \mathcal{L}_{\text{fus}}^{\mathbf{S}}}{\partial \phi \, \partial \theta_s}\right)^\top \nabla_{\theta_s'} \mathcal{L}_{\text{temp}}^{\mathbf{Q}}.$$

13: Update the dynamic loss generator:

$$\phi \leftarrow \phi - \gamma \nabla_\phi \mathcal{L}_{\text{temp}}^{\mathbf{Q}}.$$

14: **end if**
15: Freeze g_ϕ and sample a standard training clip $\mathcal{B}$.
16: Generate the learned dynamic kernels and fused result:

$$\mathcal{K}_\phi(\mathcal{B}) = \left\{K_\phi^L(\mathcal{B}), K_\phi^H(\mathcal{B})\right\}, \qquad F_\mathcal{B} = f_\theta(\mathcal{B}).$$

17: Compute the fusion objective:

$$\mathcal{L}_{\text{fus}}^\mathcal{B} = \mathcal{L}_{\text{fus}}\left(F_\mathcal{B}; \mathcal{K}_\phi(\mathcal{B})\right).$$

18: Update the fusion network:

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{fus}}^\mathcal{B}.$$

19: **end for**
20: **return** f_θ

A.2 Implementation Details of MIRA

Given the infrared and visible feature maps, MIRA first applies modality-specific normalization and independent 1×1 convolutions to obtain key and value representations. The features are partitioned into non-overlapping local windows. Within each window, the value features are spatially averaged and concatenated to condition a set of learnable anchor tokens:

$$\mathbf{A} = \mathbf{A}_0 + \text{Broadcast}\left(f_{\text{ctx}}\left([\text{Avg}(\mathbf{V}_{\text{IR}}), \text{Avg}(\mathbf{V}_{\text{VI}})]\right)\right). \tag{18}$$

Here, $\mathbf{A}_0$ denotes the learnable anchor-token matrix, $\mathbf{A}$ denotes the context-conditioned anchor matrix, and $\mathbf{V}_m$ denotes the window-wise value matrix of modality $m \in \{\text{IR}, \text{VI}\}$. Avg performs spatial averaging within a window, $[\cdot, \cdot]$ denotes feature concatenation, and f_{ctx} is a two-layer learnable projection with GELU activation. Broadcast replicates the projected window context across all anchor tokens in that window.

The conditioned anchors are normalized and projected into queries, which independently retrieve preliminary observations from both modalities. A learnable refinement network processes these observations and their absolute difference to update the anchors. The refined queries then retrieve the final modality observations $\mathbf{O}_{\text{IR}}$ and $\mathbf{O}_{\text{VI}}$ using local cross-attention. The attention employs bidirectional softmax normalization, shared relative-position biases, and learnable null tokens for rejecting unsupported matches.

For each modality, matching evidence is computed as the product of the non-null matching response and attention concentration. A learnable estimator then predicts relative modality weights from both observations, their absolute difference, and the matching evidence. The predicted logits are adjusted by the logarithm of the corresponding evidence and normalized using softmax.

Following Eq. (14), these weights construct the fusion reference $\mathbf{C}$. We further compute the modality residuals as

$$\mathbf{U}_m = \psi_m\left(\tanh\left(\frac{\mathbf{O}_m - \mathbf{C}}{\sigma_{\text{res}}}\right)\right), \quad m \in \{\text{IR}, \text{VI}\}, \tag{19}$$

$$\sigma_{\text{res}} = \sqrt{\text{Mean}_{\text{ch}}\left(|\mathbf{O}_{\text{IR}} - \mathbf{O}_{\text{VI}}|^2\right) + \epsilon}. \tag{20}$$

Here, ψ_m denotes the modality-specific 1×1 projection, σ_{res} is the residual normalization scale, Mean_{ch} denotes channel-wise averaging, and ϵ is a small positive constant for numerical stability. The hyperbolic tangent limits the normalized residual magnitude.

Finally, a learnable routing network predicts the three normalized routing weights from the fusion reference, absolute observation difference, and modality matching evidence. Following Eq. (15), the weighted reference and residual representations are combined, followed by a 1×1 output projection and feature refinement to produce the final fused representation.

B SUPPLEMENTARY TABLES

Analysis of Cross-Modal Fusion. Table A1 presents the learned cross-modal fusion statistics at high, medium, and low resolutions.

The first two columns report the relative modality weights assigned to infrared (IR) and visible (VI) observations when constructing the fusion reference. The remaining three columns report the routing weights assigned to the fusion reference, IR residual, and VI residual, respectively.

The results reveal a clear resolution-dependent fusion strategy. At high and medium resolutions, the fusion reference is predominantly constructed from VI observations, with VI weights of 98.05% and 82.77%, respectively. In contrast, the low-resolution representation assigns a substantially higher weight to IR observations (85.94%).

A similar resolution-dependent preference is observed in the residual routing. The VI residual receives higher routing weights at high and medium resolutions, whereas the IR residual receives a higher weight at low resolution. Meanwhile, the fusion-reference branch maintains substantial routing weights across all three resolutions (44.74%–55.78%).

These observations indicate that the model learns resolution-dependent modality weighting and residual routing rather than applying a uniform fusion strategy across resolutions.

Table A1: Resolution-dependent cross-modal fusion statistics. The first stage determines the modality weights for constructing the fusion reference, while the second stage controls the contributions of the fusion reference and two modality-residual branches. All values are reported as percentages.

	Stage 1: Modality Weighting		Stage 2: Fusion Routing		
Resolution	IR	VI	Fusion Ref.	IR Resid.	VI Resid.
High	1.95	**98.05**	44.74	22.08	33.18
Medium	17.23	**82.77**	45.79	13.52	40.68
Low	**85.94**	14.06	55.78	37.16	7.06

Table A2: Quantitative comparison on VTMOT. Baselines are ordered by publication year. The best and second-best results are highlighted in red and purple, respectively.

Methods	EN↑	MI↑	SSIM↑	VIF↑	flowD↓	BiSWE↓
U2Fusion (Xu et al., 2022)	6.526	2.958	0.660	0.343	14.825	8.177
LRRNet (Li et al., 2023)	6.580	2.786	0.659	0.339	12.865	9.094
DDFM (Zhao et al., 2023b)	6.757	3.040	0.708	0.371	12.396	8.164
TC-MOA (Zhu et al., 2024)	**7.150**	3.544	0.710	0.394	12.085	6.434
TIMFusion (Liu et al., 2024)	6.988	4.382	0.701	0.368	11.466	8.223
GIFNet (Cheng et al., 2025)	7.016	2.856	0.673	0.299	14.048	8.049
TemCoCo (Gong et al., 2025)	7.031	3.616	0.706	0.366	10.913	5.908
VideoFusion (Tang et al., 2026)	7.037	4.518	0.723	0.420	10.861	5.618
SDMFusion (Liu et al., 2026)	7.004	2.774	0.671	0.357	10.206	5.933
DRFusion (Li et al., 2026b)	6.467	2.664	**0.745**	0.403	10.395	7.188
MAVFusion (Li et al., 2026a)	7.052	4.180	0.702	0.423	11.338	5.591
Ours	7.113	**4.683**	0.737	**0.451**	**10.019**	**5.225**

C IMPACT OF ABNORMAL IR FRAMES IN HDO

When conducting experiments on the HDO dataset, we observed that some video sequences in the infrared (IR) mode of this dataset contained abnormal frames. Since this paper not only focuses on the fusion quality at the single-frame level but also evaluates the stability and consistency of the fusion results between consecutive frames using temporal correlation indicators, these abnormal frames may cause additional interference to the measurement of temporal indicators. Moreover, considering that the abnormal infrared input may further affect the output of the fusion model, its influence may

also be reflected in the conventional fusion quality indicators. Based on this observation, we further investigated whether the abnormal IR frames in the HDO dataset would have a significant impact on the quantitative evaluation of temporal performance and overall fusion performance.

To this end, we designed a set of additional control experiments. Specifically, while maintaining the model, inference process, and evaluation protocol exactly the same, we constructed two experimental settings. The first setting directly evaluated the complete HDO dataset, maintaining consistency with the standard experimental setup of this paper; the second setting removed the detected abnormal IR frames from the evaluation sequence and recalculated the corresponding temporal indicators and fusion quality indicators on the remaining valid frames. By comparing the quantitative results under the two settings, we aimed to isolate the factor of abnormal frames and analyze their potential impact on temporal evaluation and fusion performance evaluation. The specific experimental results and examples of abnormal frames are as follows.

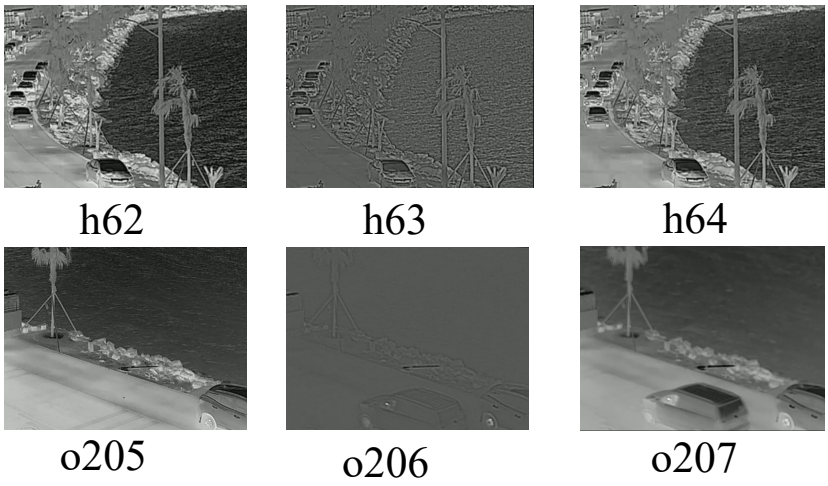


Figure A1: Examples of abnormal infrared frames in the HDO dataset.

Table A3: Influence of anomalous IR frames on temporal evaluation over the HDO dataset. Lower values indicate better temporal consistency.

Evaluation Setting	**FlowD** ↓	**BiSWE** ↓
Full HDO	4.691	2.916
w/o Anomalous Frames	**4.673**	**2.910**

Comparing the evaluation results on the full HDO dataset with those obtained after removing the anomalous IR frames, we observe that all fusion-quality metrics remain unchanged to three decimal places, while the temporal-consistency metrics exhibit only marginal improvements. These results suggest that the anomalous IR frames have a minor influence on temporal evaluation but do not noticeably affect the overall assessment of fusion quality. Therefore, to ensure a consistent and fair evaluation protocol, all results reported on HDO in this paper are computed using the complete original dataset without manually excluding any anomalous frames.