Degradation-Robust and Temporally Consistent Infrared–Visible Video Fusion via One-step Diffusion Framework

Songcheng Du[1*], Haoyuan Xu[2*], Xingyuan Li[3], Yang Zou[1], Jinyuan Liu[2], Yunpeng Bai[1†], and Ying Li[1†]

[1] Northwest Polytechnical University, Xi'an, China
[2] Dalian University of Technology, Dalian, China
[3] Zhejiang University, Hangzhou, China
dusongcheng@mail.nwpu.edu.cn, lybyp@nwpu.edu.cn

Abstract. Infrared–visible video fusion aims to integrate complementary thermal and visual information to produce stable and informative fused videos for real-world applications. While recent advances have achieved remarkable progress in image-level fusion, extending these methods to dynamic video scenarios remains challenging. Existing video fusion approaches typically adopt a 'temporal-first' processing strategy, where temporal modeling is performed within each modality before cross-modal fusion. However, such designs may become vulnerable to modality-specific degradations, including low illumination in visible videos and noise or flicker in infrared sequences, which often lead to unreliable motion estimation and temporally inconsistent fusion results. To address this limitation, we advocate a 'fusion-first' strategy and propose DRT-VF, a unified degradation-robust and temporally consistent infrared–visible video fusion method built upon a one-step diffusion architecture with a two-stage progressive training strategy. In Stage I, a Modality-Collaborative Attention module integrates complementary infrared and visible cues to establish robust intra-frame fusion representations. In Stage II, temporal consistency is modeled through a hierarchical design consisting of a Global Temporal Consistency module and a Local Motion Compensation module, which jointly enhance global temporal consistency and refine motion-sensitive regions. Extensive experiments demonstrate that DRT-VF improves temporal stability while preserving fine structural details.

Keywords: Infrared–Visible Video Fusion · Multimodal Image Fusion · Low-Level Vision

*These authors contributed equally.
†Corresponding author.

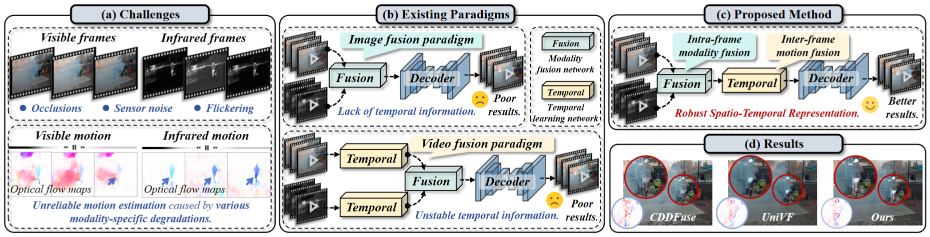

Fig. 1: Motivation and paradigm comparison. (a) **Challenges.** Modality-specific degradations, such as occlusion in visible frames and flicker in infrared sequences, lead to unreliable motion estimation and temporal artifacts, evidenced by inconsistent optical flow maps [45] in target regions. (b) **Existing Paradigms.** Image fusion methods process frames independently and ignore temporal dependencies. Many existing video fusion frameworks perform temporal modeling within each modality prior to fusion, which may propagate errors when the input observations are degraded. (c) **Proposed Method.** We instead perform cross-modal fusion before temporal refinement, establishing a more robust joint representation that facilitates stable temporal modeling.

1 Introduction

Infrared and visible modalities provide complementary cues for scene perception [30,82,86]. Visible cameras capture rich texture and color information under favorable illumination, enabling detailed representation of scene structures and appearance [59]. In contrast, infrared sensors are less sensitive to illumination variations and remain effective in low-light or adverse conditions by highlighting thermal targets and salient objects [31, 81, 87, 89]. Consequently, infrared–visible fusion has become a practical solution for various real-world applications, including surveillance [32, 38], autonomous driving, and nighttime monitoring [1, 13].

In recent years, infrared–visible image fusion has made significant progress. Various learning-based methods have been proposed to improve cross-modal information integration [49, 69], evolving from early autoencoder [85] and CNN-based frameworks [25] to recent generative and global modeling architectures [36, 38,83]. Although these methods substantially enhance spatial fusion quality, they are mainly designed for static images. In practical scenarios, infrared and visible sensors usually operate on continuous video streams [40, 80, 84]. Applying image fusion models frame by frame ignores temporal dependencies and often causes perceptual artifacts, such as temporal flicker and motion-related distortions, as shown in Fig. 1 (a), (b).

To alleviate temporal inconsistency, several video fusion methods [66, 80] introduce temporal modeling mechanisms, such as recurrent feature aggregation [44] or optical-flow-based alignment [45]. A common design in these approaches is to perform temporal modeling within each modality before cross-modal fusion [11, 80, 84] (Fig. 1 (b)). Such strategies implicitly rely on motion estimation derived from single-modality observations. However, in practical scenarios, the reliability of single-modality temporal estimation is often limited (Fig. 1 (a)). Visible sequences may suffer from occlusion, low illumination, or motion blur, while infrared frames are frequently affected by sensor noise or flickering artifacts. These modality-specific degradations can impair motion estimation accuracy [9], and the resulting errors may propagate to the subsequent fusion stage, ultimately leading to temporally unstable fusion results [29, 80].

Motivated by this observation, we rethink the conventional pipeline and advocate a "fusion-first" strategy. By integrating cross-modal complementarity before temporal restoration, the framework builds a robust joint representation that provides reliable motion cues and reduces sensitivity to modality-specific degradations, as shown in Fig. 1 (c).

Building on this insight, we propose DRT-VF (Degradation-Robust and Temporally Consistent Video Fusion), a unified infrared–visible video fusion framework based on a one-step diffusion architecture [4], implemented with rectified flow [16, 60] and trained using a two-stage progressive strategy. In Stage I, DRT-VF constructs a robust intra-frame fusion-first representation by adaptively integrating infrared and visible features through the Modality-Collaborative Attention (MCA) module. This representation is further refined in the latent space by a rectified-flow-based one-step diffusion backbone, which exploits diffusion priors to recover high-fidelity details from degraded cross-modal inputs while enabling efficient single-step inference. This stage provides a reliable foundation for subsequent temporal modeling. In Stage II, DRT-VF improves temporal consistency by separately modeling global and local temporal dynamics. Specifically, the Global Temporal Consistency (GTC) module regularizes channel-wise feature statistics across adjacent frames to suppress large-scale photometric flicker without explicit spatial alignment. Complementarily, the Local Motion Compensation (LMC) module is introduced near the final decoder stage to correct motion-induced geometric misalignment and boundary distortions at full spatial resolution [72]. By combining global temporal regularization with local motion-aware refinement, DRT-VF generates temporally stable fused videos [6] while preserving fine spatial details. Our contributions are threefold:

- We propose DRT-VF, a unified one-step diffusion-based infrared–visible video fusion framework that performs cross-modal fusion before temporal consistency modeling, alleviating the impact of modality-specific degradations on motion estimation and improving temporal robustness.
- We introduce a hierarchical temporal consistency strategy that decomposes temporal stability into global and local levels, consisting of a global temporal consistency module for suppressing flicker and a local motion compensation module for preserving motion-consistent structural details.
- Extensive experiments validate the superiority of the proposed DRT-VF in both spatial fidelity and temporal coherence across multiple benchmarks.

2 Related Work

2.1 Infrared and Visible Image Fusion

Infrared and visible image fusion aims to integrate thermal targets with rich visual textures [64, 91] for enhanced scene perception. Early deep learning approaches primarily adopted autoencoder [85] or CNN-based architectures [25] to learn modality-specific representations and perform end-to-end fusion [68]. While effective in local feature extraction, these methods are inherently limited in modeling long-range dependencies and global contextual interactions [27, 56]. To address this issue, transformer-based frameworks leverage self-attention to capture cross-modal interactions at a global scale [43, 52], whereas state-space models such as Mamba provide a computationally efficient alternative for long-context fusion [67]. Moreover, hybrid designs further combine convolutional inductive biases with global modeling modules to balance structural integrity and texture preservation [26, 81, 90]. More recently, generative modeling [14] has emerged as a dominant paradigm. Specifically, diffusion-based methods [3, 73] exploit strong generative priors to align heterogeneous infrared and visible distributions [60], demonstrating improved robustness and flexibility compared to deterministic fusion models [20,33,71,83]. In addition, semantic-aware fusion strategies [19,28,46] have also been incorporated as high-level guidance through text prompts or vision–language models, enabling controllable and semantically consistent fusion results [34, 37, 58, 69, 74].

2.2 Infrared and Visible Video Fusion

To address the limitations of static models, infrared and visible video fusion extends image-based techniques to dynamic scenarios by treating video streams as unified spatio-temporal entities, aiming to preserve cross-modal complementarity while maintaining temporal consistency across frames [12, 15, 22]. Recent research has actively explored dedicated spatio-temporal architectures to address the challenges of video fusion. TemCOCO [13] introduces a visual-semantic collaboration mechanism, explicitly utilizing high-level semantic priors to simultaneously guide cross-modal feature aggregation and inter-frame temporal consistency. To promote standardization in the field, UniVF [80] leverages temporal contexts from adjacent frames to enhance the representation capability of the current frame, while also proposing a comprehensive benchmark for systematic evaluation. Furthermore, RCVS [66], VideoFusion [51], and MambaVF [84] further investigate inter-frame temporal consistency in video fusion by employing joint registration and spatio-temporal collaborative networks, respectively.

Despite their effectiveness, most methods rely on a 'temporal-first' paradigm, assuming that single modalities can provide reliable motion cues. However, when faced with modality-specific degradations [8, 10] such as low illumination, single-modality motion estimation often fails. These temporal errors inevitably propagate to the final stage, leading to unstable fusion results.

3 Method

3.1 Preliminary

Latent Diffusion Model [47] typically reconstructs a clean latent code z_0 from Gaussian noise $z_1 \sim \mathcal{N}(0, I)$ through a stochastic, multi-step reverse process, which is computationally expensive for real-time video fusion tasks. To enable efficient generation, we adopt Rectified Flow (RF) [42, 60], a flow-based generative framework that learns a deterministic transport trajectory between noise and data distributions via an ordinary differential equation.

Given a clean latent representation z_0 and a noise sample z_1, the Rectified Flow formulation defines a linear interpolation path:

$$z_t = tz_1 + (1-t)z_0, t \in [0,1]. \tag{1}$$

The generative process is governed by an Ordinary Differential Equation (ODE) $\frac{dz_t}{dt} = v_\theta(z_t, t, c)$, where the network v_θ is optimized to predict the target velocity vector $(z_1 - z_0)$ pointing from data to noise:

$$\mathcal{L}_{RF} = \mathbb{E}_{z_0,z_1,t}\left[\|v_\theta(z_t,t,c) - (z_1 - z_0)\|^2\right]. \tag{2}$$

By learning this straight-line trajectory, the model enables ultra-fast inference. During the testing phase, the clean latent representation can be deterministically generated in a single step using the Euler method:

$$\hat{z}_0 = z_1 - v_\theta(z_1, t=1, c). \tag{3}$$

where c represents the cross-modal conditional features provided by our fusion network. This one-step rectified-flow paradigm dramatically accelerates inference while preserving strong generative priors necessary for robust image synthesis.

3.2 Framework Overview

Given a synchronized infrared–visible video sequence, DRT-VF aims to produce a temporally stable and perceptually consistent fused video by progressively organizing cross-modal fusion and temporal consistency modeling within a unified one-step diffusion framework. The overall design performs fusion before temporal refinement. Stage I first constructs a degradation-robust intra-frame fusion representation, which serves as a reliable spatial condition for the one-step diffusion backbone. Stage II then builds upon this representation and introduces hierarchical temporal modeling to suppress global flicker and correct local motion-induced inconsistencies. In this way, the two stages form a coherent progressive pipeline: robust cross-modal representation learning provides the foundation for stable temporal refinement.

Stage I: Intra-frame Cross-modal Integration. In the first stage, DRT-VF focuses on constructing a robust fusion-first representation for each frame before temporal modeling. Infrared and visible features are adaptively integrated through the Modality-Collaborative Attention (MCA) module, and the resulting representation is further refined in the latent space by a rectified-flow-based one-step diffusion backbone. This stage exploits diffusion priors to recover high-fidelity details from degraded cross-modal inputs while enabling efficient single-step inference, thereby providing a reliable spatial foundation for subsequent temporal consistency modeling.

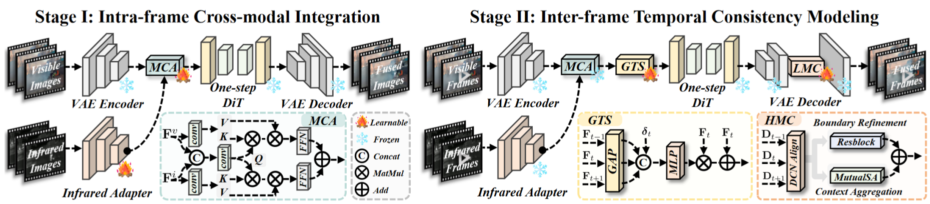

Fig. 2: Overview of proposed DRT-VF framework. Stage I focuses on establishing a robust spatial representation in image stream. Modality-specific features are integrated by MCA module to form a unified semantic anchor, which a one-step DiT further refines. Stage II ensures inter-frame consistency via a GTC module for global flicker suppression and an LMC module for local motion restoration.

Specifically, we first adopt a pre-trained variational autoencoder (VAE) [60] to encode visible images into a compact latent representation, which preserves rich structural and textural cues. In parallel, a lightweight and trainable infrared adapter [65] is employed to extract modality-aware features from infrared images, enabling flexible adaptation to thermal characteristics and noise patterns [63]. These two modality-specific representations are subsequently fed into a Modality-Collaborative Attention (MCA) module, which dynamically integrates thermal saliency from the infrared modality with fine-grained structural details from the visible modality.

Further, the unified intra-frame representation is passed to a one-step Diffusion Transformer (DiT), which serves as a generative refinement backbone. Building upon the efficient rectified flow paradigm [60], the DiT leverages strong generative priors [55]. It directly maps the fused representations to a high-fidelity latent manifold [17] in a single step. The refined latent is then decoded by a pre-trained VAE decoder [60] to produce the final image-level fusion result.

Stage II: Inter-frame Temporal Consistency Modeling. Building upon the degradation-robust fusion representation established in Stage I, Stage II introduces hierarchical temporal consistency modeling to stabilize the fused video across frames [53]. Specifically, DRT-VF decomposes temporal consistency into two complementary levels: global response stabilization and local motion-aware refinement. First, prior to the diffusion conditioning, a Global Temporal Consistency (GTC) module operates directly on the unified latent representations produced by the MCA module. Working in this low-resolution latent space, GTC regularizes channel-wise feature statistics to suppress global photometric fluctuations, without requiring explicit spatial alignment across frames while avoiding interference with high-frequency structural details [76].

Subsequently, to handle motion-induced local inconsistencies, a Local Motion Compensation (LMC) module is embedded immediately before the decoder's final layer. Operating at full spatial resolution, LMC focuses on correcting geometric misalignment and boundary distortion in dynamic regions through a dual-branch structure. By combining global latent regularization with local high-resolution refinement, the proposed design maintains stable temporal behavior while preserving fine spatial details.

3.3 Modality-Collaborative Attention Module

The Modality-Collaborative Attention (MCA) module is designed to effectively preserve and integrate complementary information from the infrared and visible modalities while suppressing modality-specific degradations. Rather than enforcing early alignment or hard fusion, the MCA module adopts a query-driven cross-attention mechanism [24] to enable flexible, content-adaptive interactions between the two visual streams.

Let $\mathbf{F}^v$ and $\mathbf{F}^i$ denote the latent features extracted from the visible and infrared images, respectively. To construct a shared fusion context, these two features are first concatenated along the channel dimension and projected through a convolutional layer to obtain a unified query representation $\mathbf{Q}$.

Instead of directly merging the features, the MCA module utilizes $\mathbf{Q}$ as a shared semantic anchor to selectively attend to each modality [54]. Specifically, the visible and infrared features are independently projected to generate modality-specific key ($\mathbf{K}$) and value ($\mathbf{V}$):

$$(\mathbf{K}^v, \mathbf{V}^v) = \phi^v(\mathbf{F}^v), \quad (\mathbf{K}^i, \mathbf{V}^i) = \phi^i(\mathbf{F}^i), \tag{4}$$

where $\phi^v(\cdot)$ and $\phi^i(\cdot)$ denote learnable linear projection operators.

Cross-attention is then performed independently for each modality to aggregate context-aware features:

$$\mathbf{A}^v = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{v\top}}{\sqrt{d}}\right)\mathbf{V}^v, \quad \mathbf{A}^i = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{i\top}}{\sqrt{d}}\right)\mathbf{V}^i. \tag{5}$$

Finally, each attended output is processed by a Feed-Forward Network (FFN) [18] before being aggregated to form the unified intra-frame representation:

$$\mathbf{F}^{fuse} = \mathrm{FFN}(\mathbf{A}^v) + \mathrm{FFN}(\mathbf{A}^i). \tag{6}$$

3.4 Global Temporal Consistency Module

Temporal flicker in multi-modal video fusion typically appears as abrupt frame-wise response fluctuations, often caused by inconsistent global activation statistics across adjacent frames, leading to perceptible brightness or contrast oscillations. To address this, we introduce a Global Temporal Consistency (GTC) module, which operates in the global semantic space to enforce channel-wise temporal stability without altering spatial structures.

Given fused features from three consecutive frames, $\{\mathbf{F}_{t-1}, \mathbf{F}_t, \mathbf{F}_{t+1}\} \in \mathbb{R}^{c\times h\times w}$, we first compress the spatial dimension via global average pooling (GAP):

$$\mathbf{u}_i = \mathrm{GAP}(\mathbf{F}_i), \quad i \in \{t-1, t, t+1\}, \tag{7}$$

where $\mathbf{u}_i \in \mathbb{R}^c$ captures channel-wise global responses.

To explicitly guide the model toward capturing frame-wise global instability, we introduce a deviation residual $\boldsymbol{\delta}_t = \mathbf{u}_t - \frac{\mathbf{u}_{t-1}+\mathbf{u}_{t+1}}{2}$, which serves as an explicit indicator of abnormal response variations associated with temporal flicker.

We therefore construct a unified temporal descriptor by concatenating the pooled statistics and the deviation term, which is then passed through a lightweight multi-layer perceptron (MLP) followed by a sigmoid activation to produce channel-wise modulation weights, $w_t = \sigma\big(\mathrm{MLP}([\mathbf{u}_{t-1}, \mathbf{u}_t, \mathbf{u}_{t+1},\ \boldsymbol{\delta}_t])\big)$, where $w_t \in \mathbb{R}^c$.

Finally, the modulation is applied to the current-frame features $\mathbf{F}_t$ in a residual manner to preserve the underlying latent manifold necessary for the subsequent generative diffusion process, $\mathbf{F}_t^{\mathrm{out}} = \mathbf{F}_t + \mathbf{F}_t \odot w_t$, where $\odot$ denotes channel-wise multiplication.

3.5 Local Motion Compensation Module

Although the preceding global temporal stabilization alleviates large-scale response fluctuations, it operates on low-resolution latent semantic statistics and is therefore insufficient to capture fine-grained temporal variations induced by object motion. To further enhance temporal consistency at the local scale, we propose the Local Motion Compensation (LMC) module, which operates at full resolution for performing fine-grained geometric alignment and enhancing temporal consistency through a decoupled dual-branch architecture.

Let $\{\mathbf{D}_{t-1}, \mathbf{D}_t, \mathbf{D}_{t+1}\} \in \mathbb{R}^{C\times H\times W}$ denote the three consecutive feature representations before the final decoder layer. To establish pixel-wise correspondence across the temporal dimension, we first employ DCN layers [88] with offset layers [13] to align the neighboring features toward the reference frame $\mathbf{D}_t$.

$$\hat{\mathbf{D}}_k = \mathrm{DCN}(\mathbf{D}_k, \mathbf{D}_t), \quad k \in \{t-1, t+1\} \tag{8}$$

where $\hat{\mathbf{D}}_k$ represents the spatially aligned features. However, directly aggregating these aligned features often introduces interpolation blur and ghosting artifacts, especially in dynamic regions where motion estimation is unreliable. To address this limitation, we propose a decoupled fusion architecture comprising a Context Aggregation branch and a Boundary Refinement branch. The outputs of the two branches are finally summed to obtain the locally compensated feature, jointly improving temporal coherence and motion-boundary preservation.

Context Aggregation Branch. This branch is designed to aggregate spatio-temporal [39] semantic contexts while filtering out misaligned pixels. We compute the Mutual Self-Attention (MutualSA) [2,57] between the center frame and the aligned temporal neighbors. By treating $\mathbf{D}_t$ as the *Query* and the aligned features $\hat{\mathbf{D}}_k$ as *Keys* and *Values*, the module adaptively aggregates complementary temporal information [70]:

$$\mathbf{D}_t^{MSA} = \mathrm{MutualSA}(\mathbf{D}_t, \hat{\mathbf{D}}_{t-1}) + \mathrm{MutualSA}(\mathbf{D}_t, \hat{\mathbf{D}}_{t+1}) \tag{9}$$

where $\mathrm{MutualSA}(\cdot)$ facilitates cross-frame interaction to reinforce the temporal consistency of local textures and dynamic details. The output of the Context

Aggregation branch is obtained by:

$$\mathbf{D}_t^{CA} = \mathbf{D}_t + \mathrm{Conv}(\mathbf{D}_t^{MSA}) \tag{10}$$

Boundary Refinement Branch. In parallel, the Boundary Refinement branch is designed to compensate for motion-related structural degradation by explicitly modeling discrepancies in the gradient domain. We compute the spatial gradient $\nabla(\cdot)$ to capture high-frequency contours [7, 62]. The temporal gradient residual is formulated as the absolute difference:

$$\mathcal{G}_t = |\nabla\hat{\mathbf{D}}_{t-1} - \nabla\mathbf{D}_t| + |\nabla\hat{\mathbf{D}}_{t+1} - \nabla\mathbf{D}_t|. \tag{11}$$

By computing the absolute gradient difference, $\mathcal{G}_t$ isolates high-frequency moving contours and alignment errors from the static background. It is then fed into a Residual Block (ResBlock) to produce a boundary calibration term, and the branch output is obtained through a residual connection:

$$\mathbf{D}_{BR} = \mathbf{D}_t + \mathrm{ResBlock}(\mathcal{G}_t) \tag{12}$$

This explicit boundary modeling restores sharp edges and suppresses motion-induced blurring artifacts without amplifying noise in static regions [77, 78].

4 Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the proposed DRT-VF framework. We first detail the experimental settings, followed by quantitative and qualitative comparisons against state-of-the-art methods. Subsequently, we analyze the temporal consistency and ablation studies to validate our design choices. Finally, we demonstrate the practical utility of our method in downstream object tracking tasks.

4.1 Experimental Setup

Datasets and Metrics. We evaluate DRT-VF on four benchmark datasets for infrared–visible video fusion: HDO [66], M3SVD [51], NOT-156 [48], and the VTMOT subset processed by VFBench [80]. For clarity, VTMOT in our experiments denotes the VFBench-processed VTMOT subset rather than the raw VTMOT dataset. We select 17 videos from each dataset for testing and use the remaining videos for training. Each sequence contains approximately 100–1000 frames. Considering the computational cost of long-video evaluation, we use the first 50 frames of each test sequence for fair and consistent comparison.

These datasets cover diverse environments and motion dynamics, with several of them naturally exhibiting these degradation factors. Specifically, NOT-156 provides nighttime scenes with low-light visible frames, HDO contains significant infrared flicker, and VTMOT includes fast-motion blur, corresponding to low illumination, flicker, and motion blur, respectively. For fusion quality assessment, we adopt Mutual Information (MI), Visual Information Fidelity (VIF),

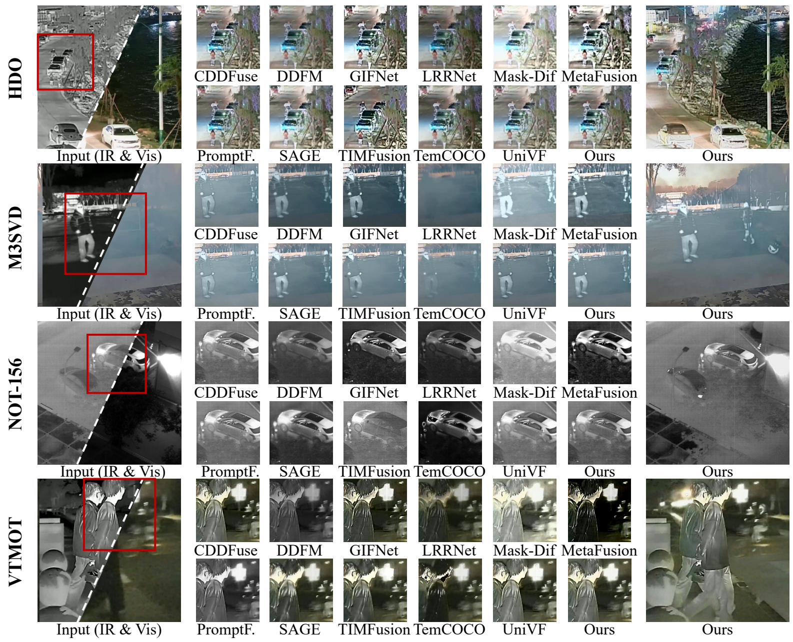

Fig. 3: Qualitative comparisons between our method and existing fusion approaches on four datasets: HDO, M3SVD, NOT-156, and VTMOT.

and Structural Similarity (SSIM). For temporal stability evaluation, we use Bidirectional Short-term Weighted Error (BiSWE) and Mean Squared Sum of Reciprocal Flow (MS2R), where lower values indicate better temporal coherence.

Implementation Details. The proposed framework is trained in a two-stage manner. In Stage I, the model is trained for 100 epochs with a batch size of 8. In Stage II, we train for 100 epochs with a batch size of 1. The optimization is performed using the AdamW optimizer with a learning rate of 1×10^{-4}. All experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU.

For the training objectives, we follow commonly adopted optimization strategies in infrared-visible fusion. In Stage I, the model is trained with a fusion-oriented objective following [69], which encourages the fused results to preserve complementary intensity, structural, gradient, and color information from the source modalities. In Stage II, we retain the same fusion objective and further incorporate a temporal consistency regularization term following [13] to encourage coherent inter-frame variations in the fused video.

4.2 Comparison with State-of-the-Art Methods

We compare our method with 12 SOTA fusion approaches, including image-based methods, i.e., CDDFuse [81], DCEVO [38], DDFM [83], GIFNet [5], LR-

Table 1: Quantitative comparison of fusion quality on four datasets.

Method	HDO			M3SVD			NOT-156			VTMOT		
	MI↑	VIF↑	SSIM↑	MI↑	VIF↑	SSIM↑	MI↑	VIF↑	SSIM↑	MI↑	VIF↑	SSIM↑
CDDFuse [81]	2.976	0.743	0.670	3.246	0.926	0.688	4.342	0.893	0.610	3.207	0.901	0.612
DCEVO [38]	3.409	0.781	0.671	**3.421**	0.931	0.691	4.219	0.884	0.641	3.822	0.953	0.654
DDFM [83]	2.783	0.695	**0.694**	2.612	0.826	**0.722**	3.479	0.791	0.640	2.822	0.778	0.648
GIFNet [5]	2.298	0.590	0.624	2.185	0.684	0.634	2.736	0.598	0.625	2.493	0.758	0.572
LRRNet [26]	2.561	0.548	0.616	2.168	0.589	0.623	3.646	0.490	0.550	2.598	0.622	0.561
Mask-Dif [50]	2.431	0.703	0.563	2.317	0.830	0.577	2.590	0.820	0.460	2.588	0.962	0.514
MetaFusion [79]	1.935	0.652	0.521	2.066	0.757	0.537	2.318	0.738	0.554	2.594	0.965	0.517
PromptF. [37]	2.778	0.713	0.665	3.046	0.895	0.689	3.647	0.830	0.586	4.172	0.970	0.654
SAGE [61]	2.599	0.621	0.649	2.485	0.681	0.666	3.763	0.798	0.640	2.699	0.745	0.607
TIMFusion [41]	3.418	0.653	0.628	3.186	0.780	0.673	3.720	0.688	0.622	3.349	0.675	0.569
TemCOCO [13]	3.250	0.647	0.568	2.218	0.466	0.441	3.713	0.703	0.605	2.814	0.712	0.615
UniVF [80]	3.483	**0.806**	0.671	3.366	0.931	0.688	4.410	0.926	0.628	4.212	0.977	0.656
Ours	**3.521**	0.792	0.680	3.373	**0.934**	0.712	**4.449**	**0.934**	**0.642**	**4.232**	**0.981**	**0.662**

RNet [26], Mask-Dif [50], MetaFusion [79], PromptFusion [37], SAGE [61], and TIMFusion [41], and video-based methods, i.e., TemCOCO [13] and UniVF [80].

Quantitative Results. As shown in Table 1, DRT-VF achieves superior performance across all benchmarks, consistently ranking within the top two in MI, VIF, and SSIM on all datasets. This demonstrates its strong ability to preserve mutual information and structural details from source scenes. For temporal stability, DRT-VF obtains the best BiSWE and MS2R scores on HDO, M3SVD, and NOT-156. On VTMOT, it achieves the best MS2R and remains competitive in BiSWE, indicating strong temporal coherence under fast-motion scenarios.

Qualitative Results. As illustrated in Fig. 3, DRT-VF exhibits superior spatial restoration fidelity, effectively mitigating ghosting and blurring artifacts observed in competing methods. As shown in the zoomed-in patches, our method preserves sharp thermal object boundaries, such as pedestrians in M3SVD and VTMOT, while retaining fine visible-background textures. This balance produces fused images that are more natural and informative.

To evaluate temporal stability, Fig. 4 (a) presents temporal profiles following the protocol of [51]. Compared with image-level fusion schemes that suffer from noticeable flicker, particularly Mask-Dif, our method maintains stronger cross-frame continuity. This is further supported by the inter-frame difference maps in Fig. 4 (b). Unlike existing approaches with scattered noise and background fluctuations, DRT-VF yields cleaner background responses and more coherent motion trajectories, confirming its effectiveness in reducing temporal artifacts.

4.3 Ablation Studies

To thoroughly investigate the contribution of our proposed modules and validate the network design, we conduct comprehensive ablation studies on the M3SVD and NOT-156 datasets. The experiments are evaluated from four perspectives: 'fusion-first' architectural paradigm, intra-frame spatial integration, inter-frame

Table 2: Quantitative comparison of temporal consistency on four datasets.

Method	HDO		M3SVD		NOT-156		VTMOT	
	BiSWE ↓	MS2R ↓	BiSWE ↓	MS2R ↓	BiSWE ↓	MS2R ↓	BiSWE ↓	MS2R ↓
CDDFuse [81]	7.398	0.220	8.882	0.238	9.983	0.588	6.358	0.374
DCEVO [38]	6.335	0.223	7.395	0.227	8.579	0.577	4.848	0.345
DDFM [83]	6.544	0.232	6.676	0.228	7.992	0.621	5.221	0.367
GIFNet [5]	7.257	0.220	7.657	0.235	10.239	0.600	5.447	0.347
LRRNet [26]	6.352	0.238	7.201	0.310	8.831	0.616	5.494	0.339
Mask-Dif [50]	8.203	0.211	9.877	0.228	10.300	0.611	6.089	0.343
MetaFusion [79]	12.259	0.224	12.556	0.251	15.597	0.609	8.327	0.421
PromptF. [37]	6.333	0.219	8.366	0.233	8.348	0.595	6.276	0.352
SAGE [61]	8.680	0.219	7.207	0.229	8.548	0.626	4.871	0.335
TIMFusion [41]	6.565	0.214	6.926	0.232	7.618	0.589	6.370	0.334
TemCOCO [13]	6.414	0.213	6.488	0.258	8.122	0.826	**4.593**	0.382
UniVF [80]	6.378	0.227	7.316	0.222	8.551	0.591	4.690	0.338
Ours	**6.232**	**0.210**	**6.479**	**0.222**	**7.542**	**0.574**	4.822	**0.332**

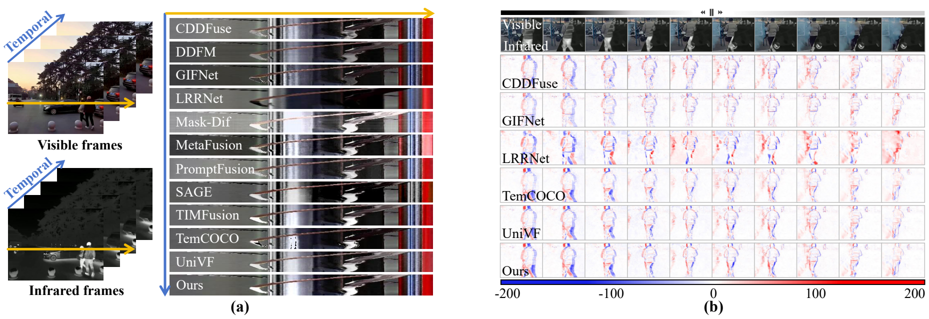

Fig. 4: Temporal consistency analysis. (a) Visual comparison via temporal profiles along the designated yellow scan lines. (b) Visualization of frame-wise differences.

temporal consistency, and the detailed design of the local motion compensation module. The results are shown in Table 3 and visualized in Fig. 5.

Effectiveness of the architectural paradigm. We first evaluate the 'fusion-first' strategy against a 'temporal-first' variant. In this variant, we keep the network components and training settings unchanged, but move temporal modeling before cross-modal fusion. Specifically, the temporal modules are applied independently to the infrared and visible feature streams, and the temporally processed modality-specific features are then fed into the MCA module and the one-step diffusion backbone for fusion. As shown in Table 3, performing temporal modeling on degraded single-modality inputs severely disrupts motion estimation, leading to significant drops in all metrics. Removing the entire Stage II (w/o Stage II) achieves high spatial scores, as enforcing temporal consistency inevitably entails a minor compromise in single-frame clarity. However, this variant completely fails to suppress flicker, resulting in the worst BiSWE and MS2R values. This confirms that explicit temporal modeling is indispensable.

Effectiveness of the components in Stage I. We analyze the components of Stage I by replacing the MCA module with simple concatenation (w/o MCA) and

Table 3: Ablation study on M3SVD and NOT-156 datasets. **Bold** and <u>underline</u> indicate best and second-best results, respectively.

Variant	M3SVD					NOT-156				
	MI↑	VIF↑	SSIM↑	BiSWE↓	MS2R↓	MI↑	VIF↑	SSIM↑	BiSWE↓	MS2R↓
Temporal-First	3.152	0.883	0.651	7.854	0.283	4.125	0.852	0.594	8.956	0.652
w/o Stage II	**3.392**	**0.941**	**0.724**	8.542	0.312	**4.466**	**0.939**	**0.653**	9.684	0.725
w/o MCA	3.214	0.892	0.686	6.551	0.237	4.253	0.887	0.616	7.625	0.591
w/o DiT	3.187	0.865	0.672	6.613	0.241	4.204	0.862	0.608	7.702	0.607
w/o GTC	3.361	0.925	0.706	8.216	<u>0.228</u>	4.432	0.921	0.636	9.458	<u>0.582</u>
w/o LMC	3.315	0.896	0.682	6.582	0.296	4.385	0.886	0.612	7.683	0.694
w/o DCN	3.332	0.907	0.693	6.514	0.286	4.406	0.902	0.622	7.605	0.685
w/o CA	3.295	0.912	0.688	6.492	0.232	4.367	0.914	0.627	7.564	0.595
w/o BR	3.344	0.916	0.698	<u>6.486</u>	0.246	4.415	0.911	0.631	<u>7.552</u>	0.612
Ours	<u>3.373</u>	<u>0.934</u>	<u>0.712</u>	**6.479**	**0.222**	<u>4.449</u>	<u>0.934</u>	<u>0.642</u>	**7.542**	**0.574**

bypassing the diffusion backbone (w/o DiT). We find that the absence of MCA leads to a drop in MI and SSIM, indicating that our query-driven mechanism is more effective at integrating multimodal information and preserving thermal saliency. Similarly, removing the DiT refinement reduces structural fidelity, which proves that the one-step diffusion backbone is crucial for recovering high-fidelity details and stabilizing the generative manifold.

Effectiveness of the components in Stage II. We further investigate Stage II, which targets temporal stability at both global and local levels. We observe that discarding the GTC module (w/o GTC) results in perceptible brightness oscillations across frames, evidenced by the increased BiSWE in Table 3 and discontinuous trajectories in Fig. 5. On the other hand, removing the LMC module (w/o LMC) primarily impacts local motion dynamics, causing severe interpolation blur and ghosting artifacts in regions with moving objects.

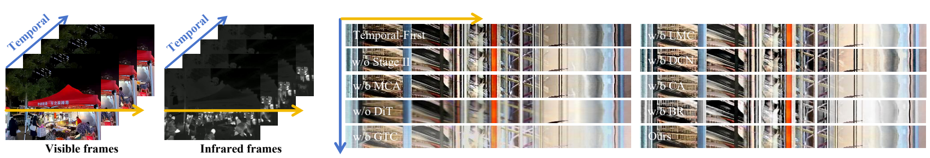

Fig. 5: Ablation study on temporal consistency. Temporal profiles along yellow scan-lines highlight contribution of each module to achieving smooth cross-frame transitions.

Detailed design of local motion compensation. To further dissect the LMC module, we study its internal dual-branch design and alignment mechanism. As shown in Fig. 5 and Table 3, removing the DCN alignment process and directly aggregating neighboring features (denoted as w/o DCN) leads to visible ghosting artifacts and a sharp increase in MS2R (from 0.574 to 0.685 on NOT-156). This

Table 4: Quantitative comparison of object tracking performance.

Method	AUC ↑	DP@20 ↑	SR@0.5 ↑	SR@0.75 ↑
CDDFuse [81]	0.220	0.209	0.208	0.084
DCEVO [38]	0.221	0.211	0.189	0.082
DDFM [83]	<u>0.240</u>	0.215	<u>0.221</u>	0.090
GIFNet [5]	0.224	0.225	0.216	0.089
LRRNet [26]	0.223	<u>0.231</u>	0.217	<u>0.108</u>
Mask-Dif [50]	0.206	0.203	0.197	0.100
MetaFusion [79]	0.231	0.228	0.205	0.099
PromptF. [37]	0.197	0.186	0.177	0.081
SAGE [61]	0.216	0.200	0.205	0.088
TIMFusion [41]	0.195	0.166	0.182	0.095
TemCOCO [13]	0.185	0.187	0.150	0.077
UniVF [80]	0.212	0.217	0.196	0.071
Ours	**0.244**	**0.236**	**0.248**	**0.116**

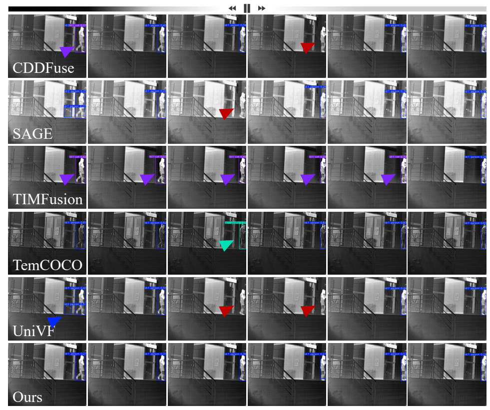

Fig. 6: Visual comparison of object tracking.

highlights the critical role of the DCN alignment process in establishing accurate pixel-wise correspondence.

W/o Context Aggregation Branch (CAB): Discarding the MutualSA mechanism weakens the aggregation of spatiotemporal contexts, leading to lower MI, VIF, and SSIM. This proves that CAB effectively leverages neighboring information to enhance the consistency and richness of local textures.

W/o Boundary Refinement Branch: Removing the gradient-based branch produces structurally aligned but blurry motion boundaries. This observation validates that explicit gradient-domain calibration is uniquely effective in restoring sharp, high-frequency contours and eliminating motion-induced artifacts in dynamic regions.

4.4 Downstream Application

To assess the practical utility of our fused videos for machine vision [35], we conduct object tracking benchmarks on the NOT-156 dataset. We employ a zero-shot configuration using ByteTrack [75] paired with a frozen, pre-trained YOLOv11n [21] detector. Quantitative evaluation is performed using three standard metrics: Area Under the Curve (AUC), Distance Precision (DP), and Success Rate (SR) [23].

As reported in Table 4, DRT-VF outperforms all baselines across all tracking metrics, indicating that our method provides superior structural clarity for accurate localization. Fig. 6 illustrates a representative challenging case where the target blends into the surroundings. Unlike existing methods that exhibit intermittent tracking failures, our approach maintains a stable trajectory and higher prediction confidence. This confirms that the joint-modality representations established by DRT-VF effectively emphasize target-background contrast, thereby offering a more discriminative representation for real-world applications.

5 Conclusion

In this paper, we presented DRT-VF, a unified framework for infrared–visible video fusion built upon a one-step diffusion architecture with a progressive two-stage training strategy. The proposed approach first establishes robust cross-modal representations through modality-collaborative fusion, and then explicitly models temporal consistency across frames. To achieve stable temporal behavior, we introduced a hierarchical temporal consistency design that combines global response regularization with local motion-aware refinement. This strategy effectively suppresses temporal flicker while preserving detailed spatial structures in dynamic scenes. Extensive experiments demonstrate that DRT-VF produces temporally stable and perceptually coherent fusion videos under challenging conditions, outperforming existing methods across multiple evaluation metrics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62271400.

References

1. Bai, H., Zhao, Z., Zhang, J., Wu, Y., Deng, L., Cui, Y., Jiang, B., Xu, S.: Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. IJCV **133**(5), 2547–2567 (2025)
2. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: ICCV. pp. 22560–22570 (2023)
3. Chen, L., Wang, J., Pan, Z., Zhu, B., Yang, X., Zhang, C.: Detail++: Training-free detail enhancer for t2i diffusion models. IEEE TIP (2026)
4. Chen, L., You, T., Liu, H., Bao, Z., Jiao, J., Han, X., Ou, Z., Sun, T., Mou, X., Jin, X., et al.: Echo: Efficient chest x-ray report generation with one-step block diffusion. arXiv preprint arXiv:2604.09450 (2026)
5. Cheng, C., Xu, T., Feng, Z., Wu, X., Tang, Z., Li, H., Zhang, Z., Atito, S., Awais, M., Kittler, J.: One model for all: Low-level task interaction is a key to task-agnostic image fusion. In: CVPR. pp. 28102–28112 (2025)
6. Cheng, Z., Zhang, X., Di, D., Wei, C., Li, H., Yang, X.: Moca: Modeling object consistency for 3d camera control in video generation. In: ICLR (2026)
7. Du, N., Wei, S., Tong, Y., Du, S., Liu, Y.: Fame-net: Frequency-adaptive mixture-of-experts network for robust infrared small target detection. IEEE TGRS (2026)
8. Du, S., Bai, Y., Ma, J., Liu, M., Li, Y.: Frequency-decoupled learning for joint thin-cloud removal and pansharpening. In: ICASSP. pp. 9282–9286. IEEE (2026)
9. Du, S., Leng, Y., Liang, X., Li, J., Liu, W., Du, Q.: Degradation aware unfolding network for spectral super-resolution. IEEE GRSL **21**, 1–5 (2024)
10. Du, S., Zou, Y., Li, J., Liu, M., Li, Y., Shang, C., Shen, Q.: Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In: AAAI. vol. 40, pp. 3696–3704 (2026)

11. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. IJCV **134**(4), 152 (2026)
12. Ellmauthaler, A., Pagliari, C.L., da Silva, E.A., Gois, J.N., Neves, S.R.: A visible-light and infrared video database for performance evaluation of video/image fusion methods. Multidimensional Systems and Signal Processing **30**(1), 119–143 (2019)
13. Gong, M., Zhang, H., Yi, X., Tang, L., Ma, J.: Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In: ICCV. pp. 14326–14335 (2025)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
15. Hu, H.M., Wu, J., Li, B., Guo, Q., Zheng, J.: An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels. IEEE TMM **19**(12), 2706–2719 (2017)
16. Huang, G., Mao, J., Huang, F., Liu, F., Luo, X., Liang, Y., Lu, J., Wang, X., Liu, P., Fu, R., Huang, R., Huang, S.L.: Exposure bias can alleviate itself via directional and frequency rectification in flow matching (2026), `https://arxiv.org/abs/2606.28226`
17. Jia, J., Liu, Y., Sun, Y., Chen, H., Qin, F., Wang, C., Peng, Y.: Geodesic prototype matching via diffusion maps for interpretable fine-grained recognition. In: ICASSP. pp. 1786–1790. IEEE (2026)
18. Jia, J., Wang, J., Liu, Y., Jing, H., Wu, Y., Wu, X., Zheng, Y.: This looks distinctly like that: Grounding interpretable recognition in stiefel geometry against neural collapse. arXiv preprint arXiv:2603.08374 (2026)
19. Jia, J., Wu, Y., Chen, H., Jing, H., Wang, H., Bu, J., Wu, L.: Unsupervised causal prototypical networks for de-biased interpretable dermoscopy diagnosis. arXiv preprint arXiv:2602.23752 (2026)
20. Jia, J., Zheng, H., Wu, Y., Chen, H., Wang, H., Bu, J., Wu, L.: Scout: Selective coupling via optimal unbalanced transport for interpretable text classification. In: ACL. pp. 6413–6431 (2026)
21. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements (2024). arXiv preprint arXiv:2410.17725 (2024)
22. Kumar, P., Mittal, A., Kumar, P.: Fusion of thermal infrared and visible spectrum video for robust surveillance. In: Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006, Madurai, India, December 13-16, 2006. Proceedings. pp. 528–539. Springer (2006)
23. Li, B., Dong, H., Zhang, D., Zhao, Z., Sun, H., Gao, J.: Exploring efficient open-vocabulary segmentation in the remote sensing. In: AAAI. vol. 40, pp. 5982–5991 (2026)
24. Li, B., Huo, T., Dong, H., Zhang, D., Zhao, Z., Gao, J., Li, X.: Towards realistic open-vocabulary remote sensing segmentation: Benchmark and baseline. arXiv preprint arXiv:2604.15652 (2026)
25. Li, H., Wu, X.J., Kittler, J.: Rfn-nest: An end-to-end residual fusion network for infrared and visible images. Information Fusion **73**, 72–86 (2021)
26. Li, H., Xu, T., Wu, X.J., Lu, J., Kittler, J.: Lrrnet: A novel representation learning guided fusion network for infrared and visible images. IEEE TPAMI **45**(9), 11040–11052 (2023)
27. Li, J., Du, S., Song, R., Li, Y., Du, Q.: Progressive spatial information-guided deep aggregation convolutional network for hyperspectral spectral super-resolution. IEEE TNNLS **36**(1), 1677–1691 (2023)

28. Li, P., Zhang, Z., Holtz, D., Yu, H., Yang, Y., Lai, Y., Song, R., Geiger, A., Zell, A.: Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. In: CVPR. pp. 40096–40107 (2026)
29. Li, X., Du, S., Zou, Y., Xu, H., Jiang, Z., Liu, J.: Unifusion: A unified image fusion framework with robust representation and source-aware preservation. arXiv preprint arXiv:2603.14214 (2026)
30. Li, X., Liu, J., Chen, Z., Zou, Y., Ma, L., Fan, X., Liu, R.: Contourlet residual for prompt learning enhanced infrared image super-resolution. In: ECCV. pp. 270–288. Springer (2024)
31. Li, X., Wang, Z., Zou, Y., Chen, Z., Ma, J., Jiang, Z., Ma, L., Liu, J.: Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In: CVPR. pp. 7534–7544 (2025)
32. Li, X., Xu, H., Li, S., Chen, X., Jiang, Z., Liu, J.: Drfusion: Drift-resilient temporally consistent infrared-visible video fusion. arXiv preprint arXiv:2605.25775 (2026)
33. Li, X., Xu, H., Zhu, X., Ma, J., Zou, Y., Jiang, Z., Liu, J.: Uncertainty-aware spatial-frequency registration and fusion for infrared and visible images. arXiv preprint arXiv:2605.13049 (2026)
34. Li, X., Zou, Y., Liu, J., Jiang, Z., Ma, L., Fan, X., Liu, R.: From text to pixels: A context-aware semantic synergy solution for infrared and visible image fusion. arXiv preprint arXiv:2401.00421 (2023)
35. Lin, H., Qin, C., Liu, Z., Pei, Q., Li, Y., Zhong, Z., Gao, X., Wang, Y., He, C., Wu, L.: Scientific image synthesis: Benchmarking, methodologies, and downstream utility. arXiv preprint arXiv:2601.17027 (2026)
36. Liu, J., Li, X., Mei, Q., Xu, H., Jiang, Z., Ma, L., Liu, R., Fan, X.: Bridging human evaluation to infrared and visible image fusion. arXiv preprint arXiv:2603.03871 (2026)
37. Liu, J., Li, X., Wang, Z., Jiang, Z., Zhong, W., Fan, W., Xu, B.: Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. IEEE/CAA Journal of Automatica Sinica **12**(3), 502–515 (2024)
38. Liu, J., Zhang, B., Mei, Q., Li, X., Zou, Y., Jiang, Z., Ma, L., Liu, R., Fan, X.: Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In: CVPR. pp. 2226–2235 (2025)
39. Liu, M., Liu, J., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H.: 4dstr: Advancing generative 4d gaussians with spatial-temporal rectification for high-quality and consistent 4d generation. In: AAAI. vol. 40, pp. 7224–7232 (2026)
40. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. arXiv preprint arXiv:2604.04198 (2026)
41. Liu, R., Liu, Z., Liu, J., Fan, X., Luo, Z.: A task-guided, implicitly-searched and meta-initialized deep model for image fusion. IEEE TPAMI **46**(10), 6594–6609 (2024)
42. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
43. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
44. Pang, S., Zeng, J., Qin, P., Qiao, G.: Ebrnet: Lightweight enhanced bidirectional recurrent network for satellite video super-resolution. IEEE TGRS (2026)
45. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR. pp. 4161–4170 (2017)

46. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
48. Sun, C., Wang, X., Fan, S., Dai, X., Wan, Y., Jiang, X., Zhu, Z., Zhong, Y.: Not-156: Night object tracking using low-light and thermal infrared: From multi-modal common-aperture camera to benchmark datasets. IEEE TGRS (2025)
49. Tang, L., Deng, Y., Yi, X., Yan, Q., Yuan, Y., Ma, J.: Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In: ACM MM. pp. 8546–8555 (2024)
50. Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. IEEE TPAMI **48**(1), 591–608 (2026)
51. Tang, L., Wang, Y., Gong, M., Li, Z., Deng, Y., Yi, X., Li, C., Xu, H., Zhang, H., Ma, J.: Videofusion: A spatio-temporal collaborative network for multi-modal video fusion. In: CVPR. pp. 19559–19569 (2026)
52. Tang, W., He, F., Liu, Y.: Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. IEEE TMM **25**, 5413–5428 (2022)
53. Wang, J., Chen, Z., Yuan, C., Li, B., Ma, W., Hu, W.: Hierarchical curriculum learning for no-reference image quality assessment. IJCV **131**(11), 3074–3093 (2023)
54. Wang, J., Duan, Y., Tao, X., Xu, M., Lu, J.: Semantic perceptual image compression with a laplacian pyramid of convolutional networks. IEEE TIP **30**, 4225–4237 (2021)
55. Wang, J., Yuan, C., Li, B., Deng, Y., Hu, W., Maybank, S.: Self-prior guided pixel adversarial networks for blind image inpainting. IEEE TPAMI **45**(10), 12377–12393 (2023)
56. Wang, W., Lu, B., Li, Y., Ji, F.: Multiscale morphology-based detection of shoreline change hotspots from aerial imagery under fluctuating water levels. Remote Sensing **18**(8), 1148 (2026)
57. Wang, Y., Turner, O., Ila, V.: Linstereo: Linear-complexity global attention for multi-scale iterative stereo matching. arXiv preprint arXiv:2606.25437 (2026)
58. Wang, Y., Miao, L., Zhou, Z., Zhang, L., Qiao, Y.: Infrared and visible image fusion with language-driven loss in clip embedding space. arXiv preprint arXiv:2402.16267 (2024)
59. Wang, Z., Zhang, J., Song, H., Ge, M., Wang, J., Duan, H.: Highlight what you want: Weakly-supervised instance-level controllable infrared-visible image fusion. In: ICCV. pp. 12637–12647 (2025)
60. Wang, Z., Zhang, J., Guan, T., Zhou, Y., Li, X., Dong, M., Liu, J.: Efficient rectified flow for image fusion. Advances in Neural Information Processing Systems **38**, 22927–22948 (2026)
61. Wu, G., Liu, H., Fu, H., Peng, Y., Liu, J., Fan, X., Liu, R.: Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In: CVPR. pp. 17882–17891 (2025)
62. Wu, Z., Wang, Y., Wen, Y., Zhang, Z., Wu, B., Tang, H.: Stereoadapter: Adapting stereo depth estimation to underwater scenes. arXiv preprint arXiv:2509.16415 (2025)
63. Xiao, X., Ma, C., Zhang, Y., Liu, C., Wang, Z., Li, Y., Zhao, L., Hu, G., Wang, T., Xu, H.: Not all directions matter: Towards structured and task-aware low-rank model adaptation. arXiv preprint arXiv:2603.14228 (2026)

64. Xiao, X., Zhang, Y., Li, X., Wang, T., Wang, X., Wei, Y., Hamm, J., Xu, M.: Visual instance-aware prompt tuning. In: ACM MM. pp. 2880–2889 (2025)
65. Xiao, X., Zhang, Y., Zhao, L., Liu, Y., Liao, X., Mai, Z., Li, X., Wang, X., Xu, H., Hamm, J., et al.: Prompt-based adaptation in large-scale vision models: A survey. arXiv preprint arXiv:2510.13219 (2025)
66. Xie, H., Sang, M., Zhang, Y., Yang, Y., Zhao, S., Zhong, J.: Rcvs: A unified registration and fusion framework for video streams. IEEE TMM (2024)
67. Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z.: Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. Visual Intelligence **2**(1), 37 (2024)
68. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE TPAMI **44**(1), 502–518 (2020)
69. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In: CVPR. pp. 27026–27035 (2024)
70. Yu, H., Jin, Y., Canevaro, A., Schmidt, J., Jordan, J., Li, P., Kaufeld, M., Lindner, S., Betz, J., Stork, W.: G2dp: Diffusion planning with spatio-temporal grid guidance. arXiv preprint arXiv:2606.26017 (2026)
71. Yue, J., Fang, L., Xia, S., Deng, Y., Ma, J.: Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. IEEE TIP **32**, 5705–5720 (2023)
72. Zhang, P., Jia, Z., Cai, Z., Weng, S., Li, S., Shi, B.: Recontraster: Making your posters stand out with regional contrast. arXiv preprint arXiv:2604.10442 (2026)
73. Zhang, P., Jia, Z., Liu, K., Weng, S., Li, S., Shi, B.: Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In: CVPR. pp. 659–669 (2026)
74. Zhang, P., Weng, S., Zhu, C., Tang, B., Jia, Z., Li, S., Shi, B.: Affective image editing: Shaping emotional factors via text descriptions. IJCV **134**(1), 16 (2026)
75. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV. pp. 1–21. Springer (2022)
76. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: CVPR. pp. 8281–8291 (2024)
77. Zhao, C., Ci, E., Xu, Y., Fan, T., Guan, S., Ge, Y., Yang, J., Tai, Y.: Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. In: NeurIPS. vol. 38, pp. 3373–3393. Curran Associates, Inc. (2025)
78. Zhao, C., Xu, Y., Chen, Z., Gu, E., Zhang, K., Liu, X., Yang, J., Tai, Y.: From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. IEEE TPAMI (2026)
79. Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H.: Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In: CVPR. pp. 13955–13965 (2023)
80. Zhao, Z., Bai, H., Ke, B., Cui, Y., Deng, L., Zhang, Y., Zhang, K., Schindler, K.: A unified solution to video fusion: From multi-frame learning to benchmarking. NeurIPS **38**, 128504–128535 (2026)
81. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: CVPR. pp. 5906–5916 (2023)
82. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Zhang, K., Xu, S., Chen, D., Timofte, R., Van Gool, L.: Equivariant multi-modality image fusion. In: CVPR. pp. 25912–25921 (2024)

83. Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: Ddfm: denoising diffusion model for multi-modality image fusion. In: ICCV. pp. 8082–8093 (2023)
84. Zhao, Z., Cui, Y., Deng, L., Bai, H., Qin, H., Feng, T., Schindler, K.: Mambavf: State space model for efficient video fusion. arXiv preprint arXiv:2602.06017 (2026)
85. Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J.: Didfuse: Deep image decomposition for infrared and visible image fusion. arXiv preprint arXiv:2003.09210 (2020)
86. Zhao, Z., Zhang, J., Bai, H., Wang, Y., Cui, Y., Deng, L., Sun, K., Zhang, C., Liu, J., Xu, S.: Deep convolutional sparse coding networks for interpretable image fusion. In: CVPR. pp. 2369–2377 (2023)
87. Zhu, P., Sun, Y., Cao, B., Hu, Q.: Task-customized mixture of adapters for general image fusion. In: CVPR. pp. 7099–7108 (2024)
88. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: CVPR. pp. 9308–9316 (2019)
89. Zou, Y., Chen, Z., Zhang, Z., Li, X., Ma, L., Liu, J., Wang, P., Zhang, Y.: Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution. IJCV **134**(1), 23 (2026)
90. Zou, Y., Ma, J., Jiao, Z., Li, X., Jiang, Z., Liu, J.: Toward real-world infrared image super-resolution: A unified autoregressive framework and benchmark dataset. In: CVPR. pp. 16365–16375 (2026)
91. Zou, Y., Zhu, X., Han, K., Ma, J., Li, X., Jiang, Z., Liu, J.: Hatir: Heat-aware diffusion for turbulent infrared video super-resolution. arXiv preprint arXiv:2601.04682 (2026)