MAVFusion: Efficient Infrared and Visible Video Fusion via Motion-Aware Sparse Interaction

Xilai Li[*,1], Weijun Jiang[*,1], Xiaosong Li[†,1], Yang Liu[1], Hongbin Wang[2], Tao Ye[3], Huafeng Li[2], and Haishu Tan[1]

[1] Foshan University, Foshan, China
20210300236@stu.fosu.edu.cn, lixiaosong@buaa.edu.cn
[2] Kunming University of Science and Technology, Kunming, China
[3] China University of Mining and Technology, Beijing, China

Abstract. Infrared and visible video fusion combines the object saliency from infrared images with the texture details from visible images to produce semantically rich fusion results. However, most existing methods are designed for static image fusion and cannot effectively handle frame-to-frame motion in videos. Current video fusion methods improve temporal consistency by introducing interactions across frames, but they often require high computational cost. To mitigate these challenges, we propose MAVFusion, an end-to-end video fusion framework featuring a motion-aware sparse interaction mechanism that enhances efficiency while maintaining superior fusion quality. Specifically, we leverage optical flow to identify dynamic regions in multi-modal sequences, adaptively allocating computationally intensive cross-modal attention to these sparse areas to capture salient transitions and facilitate inter-modal information exchange. For static background regions, a lightweight weak interaction module is employed to maintain structural and appearance integrity. By decoupling the processing of dynamic and static regions, MAVFusion simultaneously preserves temporal consistency and fine-grained details while significantly accelerating inference. Extensive experiments demonstrate that MAVFusion achieves state-of-the-art performance on multiple infrared and visible video benchmarks, achieving a speed of 14.16 FPS at 640×480 resolution. The source code is available at https://github.com/ixilai/MAVFusion.

Keywords: Video fusion · Infrared and visible · Optical flow · Motion-aware sparse interaction · Efficient video processing

1 Introduction

In visual perception tasks, a single modality often cannot fully represent all the information in a scene. Practical applications usually require combining complementary information from different modalities to overcome the limitations of any

[*] Equal contribution.
[†] Corresponding author.

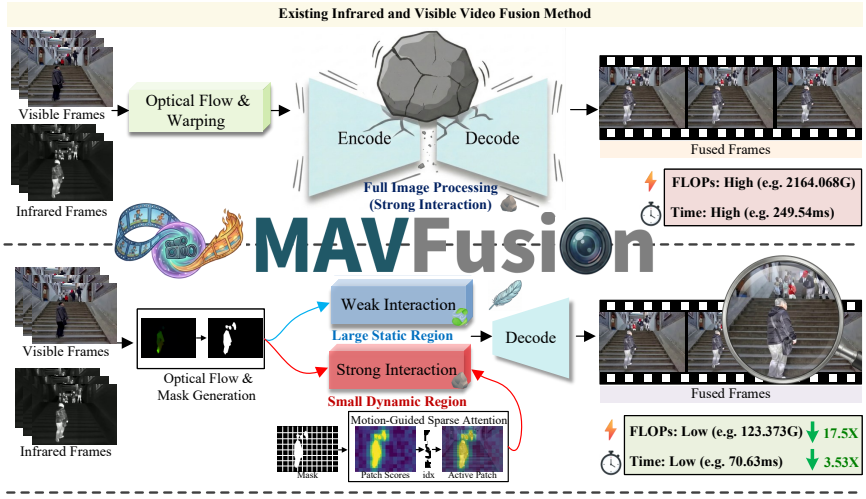

Fig. 1: Comparison of the proposed algorithm with existing video fusion method [41] in terms of multi-modal interaction strategies and computational efficiency.

single modality [11, 29, 30, 34, 39]. As an important branch of multi-modal visual fusion, infrared and visible fusion has significant practical value [2, 16, 17, 40].

Although significant progress has been made in infrared and visible fusion, most existing studies still focus on image-level tasks, such as cross-modal interaction [13, 14, 42], joint optimization with downstream tasks [5, 22, 24], and degraded-scene fusion [19, 37, 38]. These methods usually ignore temporal dependencies across frames and lack awareness of motion and dynamic changes. As a result, they may produce temporal jitter and fail to fully exploit motion cues in video scenarios. Compared with single-frame fusion, multi-modal video fusion can better capture object motion and temporal variations by leveraging inter-frame correlations. This improves the coherence of fusion results and provides useful temporal references for cross-modal interaction in dynamic scenes.

Most existing video fusion methods [27, 41] rely on global modeling to maintain temporal consistency. This approach requires spatio-temporal interaction over large areas. Although it captures dynamic relationships, it also leads to high computational cost. In addition, dense attention mechanisms usually treat all regions equally. As shown in Fig. 1, this is inefficient for long sequences, because much computation is spent on less important background areas instead of focusing on key moving objects that truly need strong fusion. Furthermore, current methods do not fully use an important advantage of video: motion information can directly indicate where fusion should focus. Motion cues, such as optical flow, can reveal dynamic changes in the scene and help locate moving regions that contain important objects, such as pedestrians and vehicles. In infrared and visible fusion, strong cross-modal interaction and precise alignment

are often most needed in these regions. In contrast, static background areas usually occupy most of the scene, and the modality with richer details (often the visible image) naturally provides better structure and texture. Applying excessive interaction in these static areas, or forcing information from a weaker or noisier modality, may disturb the original structure and semantics, leading to artifacts or unstable results. These observations suggest that video fusion requires a new framework that allocates computation according to motion cues and balances strong interaction in key dynamic regions with stable preservation of static backgrounds.

To address the above issues, we propose an efficient infrared and visible video fusion method based on motion-aware sparse attention interaction (MAVFusion). Our method focuses on two aspects: improving temporal stability through motion-based alignment, and reducing computational cost by separating the interaction of static and dynamic regions. First, we design a lightweight Motion-Aware Feature Alignment Module (MAFM) to improve cross-frame consistency and reduce ghosting. We use a frozen SEA-RAFT model [28] to estimate optical flow from visible frames for coarse alignment. Then, we predict residual optical flow in the feature space to refine the alignment and reduce errors between infrared and visible images. A motion mask is generated from the optical flow magnitude. The temporally aggregated features are injected through a motion-mask soft gate, so that motion regions benefit from multi-frame cues while static regions mainly rely on current-frame features to avoid unnecessary cross-frame mixing. Second, we propose a static–dynamic decoupled multi-modal fusion module. The fusion process is divided into a weak interaction branch for static regions and a strong interaction branch for dynamic regions. In the static branch, convolutional modules are used for local modeling, which preserves background structure and texture details in a stable and low-cost way, ensuring natural and consistent overall appearance. In the dynamic branch, guided by the motion mask, sparse attention is applied only to selected Top-K patches in key regions. Our main contributions are as follows:

- We propose an infrared and visible video fusion framework based on motion-aware sparse attention. The framework separates static and dynamic regions to allocate computation more efficiently, achieving good fusion quality with fast inference.
- We propose a lightweight motion-aware feature alignment module that enhances cross-frame consistency in motion regions while avoiding noise in static areas, effectively reducing ghosting in multi-frame fusion.
- We design a motion–guided sparse attention module that restricts complex interactions to important motion regions. This allows dynamic targets to build long-range connections with the global context, improving cross-modal fusion while reducing unnecessary computation.
- Extensive experiments on multiple benchmarks consistently show that our method outperforms state-of-the-art approaches in fusion quality, temporal stability, and computational efficiency, striking an excellent balance between performance and speed.

2 Related Work

2.1 Infrared and Visible Image Fusion

Existing infrared and visible image fusion (IVIF) methods mainly include autoencoder based and generative model-based approaches. Autoencoder-based methods [3,10,12,43] extract key features from source images and perform fusion using deep models, while generative model-based methods [9,36,44] learn latent image distributions to produce high-quality fusion results. Recent research further extends IVIF to real-world degradations, including noise [8], motion blur [4], and adverse weather [18,19]. Some methods [15,23] also exploit semantic information and task-driven modeling to improve fusion performance. However, these approaches still focus on single images and do not capture temporal information, limiting their effectiveness for video fusion.

2.2 Infrared and Visible Video Fusion

Compared to the well-studied field of static image fusion, infrared and visible video fusion must not only integrate complementary information within each frame but also maintain motion consistency and temporal stability across frames [6,7,27,41]. For example, Guo et al. [7] reduced the effect of intra-frame feature discrepancies on fusion consistency using a hierarchical fusion strategy. Zhao et al. [41] combined multi-frame learning with optical flow feature registration in a single framework, producing fusion results with better temporal coherence. Although these methods advance video fusion, they rely on global spatio-temporal modeling or dense attention mechanisms to ensure temporal consistency. This approach significantly increases computational cost and wastes processing power on unimportant background regions in long sequences.

3 Methodology

The overall structure of the proposed algorithm is shown in Fig. 2. This framework consists of four core stages: shallow feature extraction and motion estimation, Motion-Aware Feature Alignment Module (MAFM), Motion-Guided Dual-Interaction Module (MDIM), and image reconstruction decoder.

3.1 Motion-Aware Feature Alignment Module

In dynamic video fusion tasks, cross-modal spatial misalignment and object motion often cause ghosting and blur in the fusion results. To address this problem, we propose a lightweight MAFM module. This module improves alignment through cross-modal guidance and a coarse-to-fine warping strategy. The alignment process includes two stages to gradually reduce spatial differences between adjacent frames.

Multi-frame Coarse Alignment: First, to mitigate the prohibitive computational cost and large motion amplitudes at high resolutions [20, 21], we

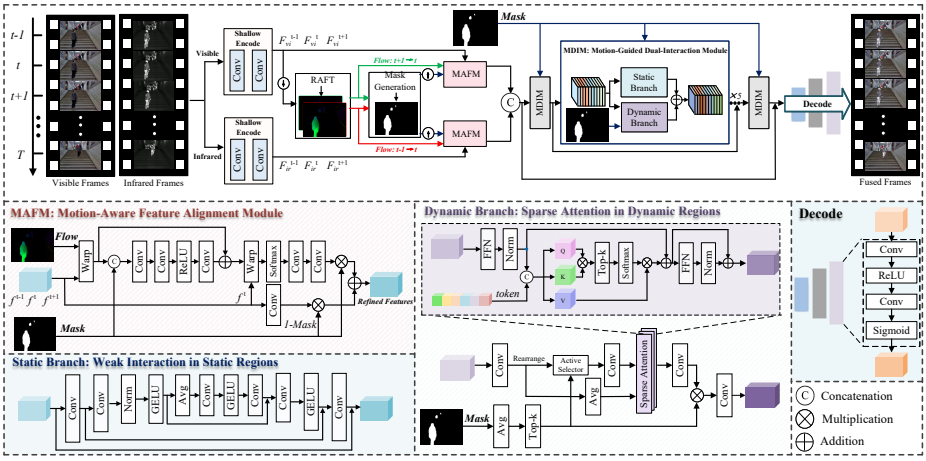

Fig. 2: The overall framework of the proposed algorithm.

estimate optical flow via SEA-RAFT [28] at a lower resolution and upsample it for initial spatial alignment. The coarse feature alignment is defined as:

$$\tilde{f}_i = \mathcal{W}(f_i, \phi_i), \quad i \in \{t-1, t, t+1\} \tag{1}$$

where $\mathcal{W}(\cdot)$ denotes the bilinear interpolation warping operator, and f_i represents the feature at time i. For adjacent frames, ϕ_{prev} and ϕ_{next} denote the optical flows from frame $t-1$ to t and from frame $t+1$ to t, respectively. For the current frame, we use their average $\phi_t = (\phi_{\text{prev}} + \phi_{\text{next}})/2$ as an initial motion estimate.

Cross-Modal Residual Refinement: To further reduce flow errors, we introduce features from the other modality as spatial anchors for cross-modal guidance. The module concatenates the anchor features, the three coarsely aligned frames, and the original flow field, and then predicts the residual flow $\Delta\phi$ using depthwise separable convolutions:

$$\Delta\phi = \mathcal{H}_{\text{refine}}\left(\left[f_{\text{anchor}}, \tilde{f}_{t-1}, \tilde{f}_t, \tilde{f}_{t+1}, \phi_{\text{prev}}, \phi_{\text{next}}\right]\right), \tag{2}$$

where $\mathcal{H}_{\text{refine}}$ represents the refinement sequence involving channel compression and spatial refinement. Finally, the refined feature $\hat{f}_t$ is generated by applying the residual compensation:

$$\hat{f}_t = \mathcal{W}\left(f_t, \phi_t + \Delta\phi\right). \tag{3}$$

After obtaining the aligned feature sequence, we use an efficient temporal aggregation strategy instead of complex 3D convolutions. By learning temporal importance weights ω_i, which are initialized to one, Softmax-normalized along the temporal dimension in the forward pass, and optimized end-to-end with the

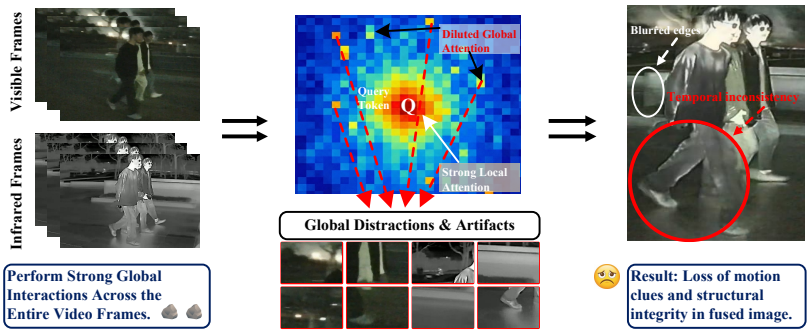

Fig. 3: Effect of Global Strong Interaction on Motion Cues and Structural Integrity.

overall training objective, MAFM adaptively estimates the contribution of each aligned frame:

$$F_{\text{agg}} = \sum_{i \in \{t-1,t,t+1\}} \text{Softmax}(\boldsymbol{\omega})_i \, \hat{f}_i. \tag{4}$$

Here, $\hat{f}_i$ denotes the aligned feature used for temporal aggregation, where the center frame uses the refined feature $\hat{f}_t$, while the adjacent frames retain the coarsely aligned features. Finally, to maintain background stability while exploiting temporal information, MAFM does not directly apply the temporally aggregated features to all regions. Instead, the motion mask is used as a soft gate to inject temporal aggregation into motion regions, while static regions mainly rely on the current-frame features. This motion-gated fusion suppresses unnecessary cross-frame mixing and preserves stable background structures, while allowing dynamic regions to benefit from temporal cues.

3.2 Motion-Guided Dual-Interaction Module

In infrared and visible fusion tasks, scenes often show clear static-dynamic differences: background regions are usually static with mostly redundant texture, while motion regions contain salient cross-modal features and important semantic targets. Existing video fusion methods apply uniform attention to all features, which leads to quadratic computational overhead. Furthermore, as shown in Fig. 3, in long-sequence, high-resolution scenarios, attention must distribute weights across many tokens, causing salient and background regions to compete and resulting in diffuse attention. This can weaken local structures and reduce temporal consistency. To address this, we propose a Motion-Guided Dual-Interaction Module. It applies weak interaction to static regions and sparse strong interaction to dynamic regions, enabling content-adaptive allocation of computation and more effective feature reconstruction.

Dynamic Branch: Sparse Attention in Dynamic Regions Dynamic regions usually contain the most critical moving targets and salient semantic infor-

mation in a scene, which require modeling long-range cross-modal dependencies. We employ an optical-flow-based motion mask $M \in \mathbb{R}^{H \times W}$ as a physical prior to guide the network to perform strong interactions only within active dynamic regions.

First, the input feature map is partitioned into non-overlapping $p \times p$ patches, resulting in a total of N patches. By applying adaptive average pooling with the same spatial scale to the mask M, we obtain a patch-level saliency score vector $S \in \mathbb{R}^N$. Based on a predefined retention ratio $\tau = 0.3$ (Top-K ratio), we set $k = \lfloor N\tau \rfloor$ and select the top-k active patches, forming an index set $\Omega \subset \{1, \ldots, N\}$, while retaining their corresponding saliency scores as soft weights W:

$$S = \text{AvgPool}_{p \times p}(M), \quad \Omega, W = \text{TopK}(S, \lfloor N \cdot \tau \rfloor). \tag{5}$$

It is worth noting that this relative ranking strategy is especially important for handling ego-motion or camera shake. Instead of a hard threshold, the Top-K selection adaptively isolates foreground objects with the strongest relative motion, keeping computation limited and preventing costly full-frame processing.

Since the identification and enhancement of dynamic targets often require reference to global contextual cues, we spatially average each patch feature to form a set of pooled global context tokens $T_{\text{global}} = \{T_i\}_{i=1}^N$:

$$T_i = \text{AvgPool}\left(X_{\text{patch}}^{(i)}\right), \quad i = 1, \ldots, N. \tag{6}$$

Subsequently, we construct an asymmetric joint sparse attention mechanism. The query vectors are generated only from the activated salient patch features X_Ω, while the keys and values are provided by the pooled global context tokens T_{global}. Since only k active patches are used as queries and all N pooled patch tokens serve as global keys and values, this design reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(k \cdot N)$, achieving an effective balance between efficiency and receptive field:

$$F_{\text{attn}} = \text{Softmax}\left(\frac{Q(X_\Omega)\, K(T_{\text{global}})^\top}{\sqrt{d}}\right) V(T_{\text{global}}). \tag{7}$$

Static Branch: Weak Interaction in Static Regions For static background regions, complex global self-attention mechanisms are often unnecessary. In the static branch, we employ lightweight depthwise separable convolutions together with standard 3×3 convolutions. This design constrains the receptive field to local neighborhoods for weak interaction, aiming to efficiently extract and preserve low-level high-frequency textures while significantly reducing computational cost.

Adaptive Reconstruction and Fusion of Dynamic and Static Features After obtaining the static texture features F_{static} and the dynamic salient features F_{attn}, they must be seamlessly fused in the spatial domain to avoid hard

boundary artifacts. We first use the saliency soft weights W to reweight the dynamic features and then apply an index-copy operation (denoted as $\mathcal{R}$) to scatter the sparse patches back to their original spatial locations. Finally, the interpolated mask M serves as a spatial gating signal to achieve adaptive reconstruction and smoothing:

$$Y = F_{\text{static}} + \text{Smooth}\big(M \odot \mathcal{R}(F_{\text{attn}} \odot W)\big). \tag{8}$$

Here, Smooth$(\cdot)$ denotes a 3×3 smoothing convolution layer, which alleviates stitching artifacts caused by multi-scale patch partitioning and ensures a natural spatial and semantic transition in the final fused feature Y.

3.3 Loss Function

To generate high-quality fused video sequences with spatial fidelity and temporal consistency, we define the total loss $\mathcal{L}_{\text{total}}$ as a weighted combination of the spatial loss $\mathcal{L}_{\text{spatial}}$ and the temporal consistency loss $\mathcal{L}_{\text{temp}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spatial}} + \gamma \mathcal{L}_{\text{temp}}, \tag{9}$$

where γ is a hyper-parameter used to balance spatial reconstruction and temporal stability in video sequences.

Spatial Fidelity Constraint The spatial loss $\mathcal{L}_{\text{spatial}}$ primarily serves to integrate complementary and salient information from the two modalities. Specifically, $\mathcal{L}_{\text{spatial}}$ consists of a pixel similarity loss [41] and a structural similarity loss [31], preserving infrared saliency and visible details in the fused frame I^F.

Temporal Consistency Constraint To mitigate inter-frame flickering and ensure smooth visual transitions in the output video, we introduce a temporal consistency loss $\mathcal{L}_{\text{temp}}$ [41]. This term explicitly enforces frame-to-frame stability by penalizing misalignments between adjacent fused frames using estimated optical flow fields $\mathcal{O}$. To avoid unreliable gradients in occluded or poorly aligned regions, we incorporate validity masks M^t_{prev} and M^t_{next} to identify reliable pixels [41]. The loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{temp}} = \mathbb{E}_{p \in M^t_{prev}} \left[\left| I^F_t(p) - \mathcal{W}(I^F_{t-1}, \mathcal{O}^F_{t-1 \to t})(p) \right|_1 \right] \\ + \mathbb{E}_{p \in M^t_{next}} \left[\left| I^F_t(p) - \mathcal{W}(I^F_{t+1}, \mathcal{O}^F_{t+1 \to t})(p) \right|_1 \right]. \end{aligned} \tag{10}$$

where p denotes the pixel coordinates, and $\mathcal{W}(\cdot, \mathcal{O})$ denotes the backward warping operation guided by the flow field $\mathcal{O}$.

4 Experiments

4.1 Experiment Settings

Datasets and Metrics To evaluate the performance of the proposed method on infrared and visible video fusion, we conduct experiments on three public video

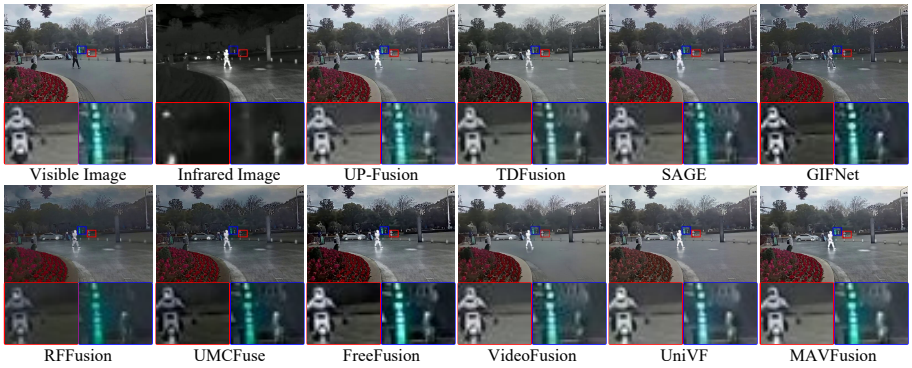

Fig. 4: Qualitative comparison of all methods on the M3SVD dataset.

datasets: M3SVD [27], HDO [35], and VTMOT [45]. The M3SVD dataset contains 30 video sequences, HDO contains 24 sequences, and VTMOT contains 90 sequences. For each dataset, we randomly select five video sequences for testing, and use the remaining sequences for training. Furthermore, to comprehensively evaluate the fusion performance, we adopt eight objective metrics, including image quality metrics and a video smoothness metric. The image quality metrics include Gradient-Based Fusion Performance (Q_G), Image Fusion Metric Based on a Multiscale Scheme (Q_M), Image Fusion Metric Based on Phase Congruency (Q_P), Piella's Metric (Q_S), Chen–Blum Metric (Q_{CB}), Visual Information Fidelity (VIF), and Normalized Weighted Performance Metric ($Q^{AB/F}$) [25]. For video smoothness evaluation, we use Motion Smoothness with Dual Reference Videos (MS2R) [41].

Comparative Methods To validate the effectiveness of the proposed method, we compare it with seven state-of-the-art image fusion methods and two video fusion methods. The image fusion methods include UP-Fusion [15], TDFusion [1], SAGE [33], GIFNet [5], RFFusion [32], UMCFuse [16], and FreeFusion [40]. The video fusion methods are VideoFusion [27] and UniVF [41]. All methods are evaluated under the same experimental settings to ensure a fair comparison.

Training Details All experiments are conducted on a machine equipped with four NVIDIA GeForce RTX 3090 GPUs. During training, input frames are randomly cropped to a size of 384×384. The batch size is set to 32 with gradient accumulation enabled. We use the Adam optimizer with an initial learning rate of 1.0×10^{-4}. After 40,000 iterations, the learning rate is exponentially decayed to 1% of its initial value.

4.2 Qualitative Comparative Experiment

Figs. 4 to 6 show a qualitative comparison of the proposed MAVFusion algorithm with competing methods on three video fusion datasets (M3SVD, HDO,

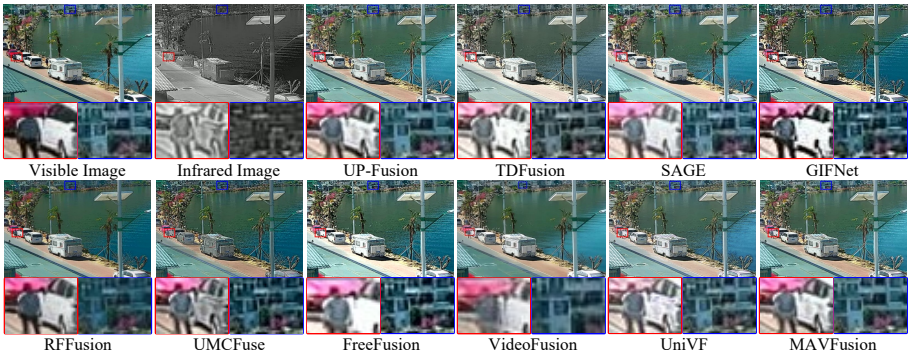

Fig. 5: Qualitative comparison of all methods on the HDO dataset.

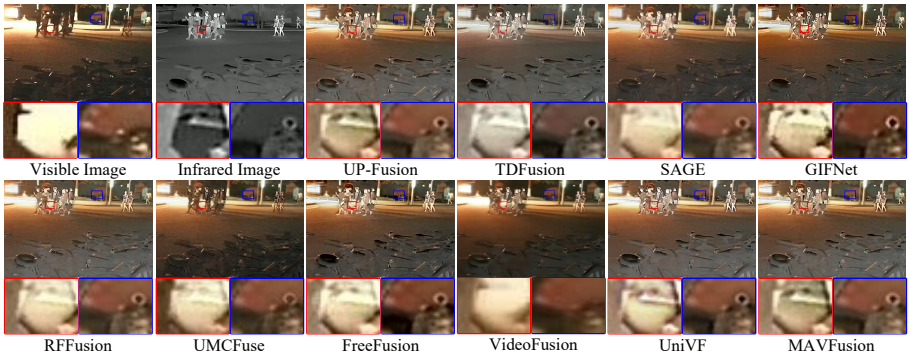

Fig. 6: Qualitative comparison of all methods on the VTMOT dataset.

VTMOT). For each result, we provide zoomed-in views of both moving and static regions. For dynamic objects, methods that ignore temporal continuity, such as UP-Fusion, SAGE, and GIFNet, often produce ghosting and blurred edges. RFFusion, UMCFuse, and FreeFusion alleviate ghosting by relying more on infrared background information, but this weakens spatial contrast and blurs scene structures. In contrast, video fusion methods such as UniVF and MAVFusion better suppress ghosting and preserve contrast. However, UniVF may over-introduce smooth infrared information into static regions, weakening visible textures and overall scene expressiveness. MAVFusion separates dynamic and static processing for region-adaptive fusion, effectively mitigating spatio-temporal conflicts in video fusion.

4.3 Quantitative Comparative Experiment

Tab. 1 reports the quantitative comparison results across three datasets, with the top two performances highlighted. The proposed algorithm achieves state-of-the-art scores across most image quality metrics. Several competitive methods,

Table 1: Quantitative comparison on M3SVD, HDO, and VTMOT datasets. **Bold** and red indicate the best performance, while green indicates the second best.

Type	Methods	Pub.	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$Q_S\uparrow$	$Q_{CB}\uparrow$	$VIF\uparrow$	$Q^{AB/F}\uparrow$	$MS2R\downarrow$
M3SVD Dataset										
Image	UP-Fusion	*AAAI 26*	0.6153	0.8126	0.5391	0.8394	0.5173	0.3492	0.6935	0.1110
	TDFusion	*CVPR 25*	0.6703	0.5606	0.5346	0.8520	0.5163	0.3382	0.6962	0.1059
	SAGE	*CVPR 25*	0.4739	0.3088	0.4433	0.8114	0.4697	0.2974	0.6110	0.1130
	GIFNet	*CVPR 25*	0.3988	0.1966	0.3318	0.7537	0.4653	0.1969	0.4823	0.1169
	RFFusion	*NeurIPS 25*	0.4371	0.2539	0.4106	0.7085	0.4071	0.3313	0.3589	0.1107
	UMCFuse	*TIP 25*	0.6256	0.5049	0.4287	0.7818	0.5015	0.2807	0.5994	0.1127
	FreeFusion	*TPAMI 25*	0.4485	0.2338	0.4027	0.6702	0.4186	0.2009	0.5438	0.1215
Video	VideoFusion	*CVPR 26*	0.3881	0.2832	0.4072	0.7964	0.4380	0.2631	0.5491	0.1593
	UniVF	*NeurIPS 25*	0.6376	0.5661	0.5529	0.8449	0.5169	0.3425	0.7049	0.1064
	MAVFusion	–	**0.6897**	**1.1544**	**0.6122**	**0.8549**	**0.5535**	**0.3814**	**0.7550**	**0.1058**
HDO Dataset										
Image	UP-Fusion	*AAAI 26*	0.5997	0.7588	0.5096	0.8441	0.5345	0.3186	0.6200	**0.9657**
	TDFusion	*CVPR 25*	0.6153	0.6787	0.4659	0.8417	0.5155	0.2981	0.6081	0.9952
	SAGE	*CVPR 25*	0.3928	0.3536	0.3643	0.7528	0.5044	0.2674	0.4620	1.0622
	GIFNet	*CVPR 25*	0.3465	0.2565	0.2987	0.6959	0.5105	0.1909	0.3823	1.0546
	RFFusion	*NeurIPS 25*	0.4478	0.3637	0.3316	0.6773	0.4645	0.2619	0.4191	1.0267
	UMCFuse	*TIP 25*	0.5879	0.6375	0.4099	0.8185	0.5140	0.2795	0.5471	1.0110
	FreeFusion	*TPAMI 25*	0.4839	0.4183	0.3667	0.6797	0.4860	0.2124	0.5044	1.0020
Video	VideoFusion	*CVPR 26*	0.3649	0.3372	0.3642	0.7391	0.4781	0.2616	0.4128	1.2651
	UniVF	*NeurIPS 25*	0.6125	0.7655	0.4953	0.8433	0.5300	0.3109	0.6364	1.0086
	MAVFusion	–	**0.6629**	**0.9996**	**0.5660**	**0.8464**	**0.5578**	**0.3359**	**0.6837**	0.9826
VTMOT Dataset										
Image	UP-Fusion	*AAAI 26*	0.5154	0.6226	0.3894	0.8486	0.3635	0.3798	0.6307	1.5948
	TDFusion	*CVPR 25*	0.5815	0.8928	0.4197	**0.8517**	0.3899	**0.4147**	0.6713	0.8346
	SAGE	*CVPR 25*	0.3285	0.3129	0.2168	0.7592	0.3819	0.2877	0.4169	0.9498
	GIFNet	*CVPR 25*	0.2963	0.2420	0.2223	0.7113	**0.4389**	0.2115	0.3834	0.9402
	RFFusion	*NeurIPS 25*	0.4899	0.3836	0.2731	0.6741	0.4138	0.2715	0.4176	0.9082
	UMCFuse	*TIP 25*	0.5968	0.9231	0.3831	0.8272	0.4077	0.3534	0.6050	0.9939
	FreeFusion	*TPAMI 25*	0.5320	0.5073	0.3705	0.7881	0.4143	0.2950	0.5749	0.9775
Video	VideoFusion	*CVPR 26*	0.2481	0.2784	0.1686	0.7448	0.3382	0.2832	0.3332	1.2241
	UniVF	*NeurIPS 25*	0.5724	0.8993	0.4276	0.8469	0.3892	0.3856	0.6705	0.8380
	MAVFusion	–	**0.6325**	**0.9741**	**0.4947**	0.8504	0.4036	0.3885	**0.7182**	**0.8301**

such as UP-Fusion, TDFusion, and UniVF, also show strong performance in image gradients, structural integrity, and perceptual quality, demonstrating their ability to preserve spatial details. Additionally, our weak-interaction design for background regions proves effective, indicating that sparse-textured backgrounds require only minimal interaction to maintain high-fidelity fusion. The strong MS2R result further demonstrates good temporal continuity. Sparse strong interaction in dynamic regions helps capture motion-aware multimodal cues, reduce ghosting, and enhance spatial contrast.

4.4 Downstream Task Experiments

To evaluate the practical potential of the proposed algorithm, we conduct downstream object detection experiments using YOLO26 [26] as the base detector.

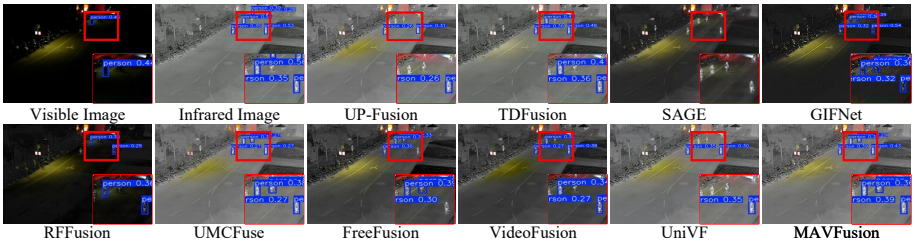

Fig. 7: Detection accuracy comparison on fused results.

Table 2: Quantitative comparison of different ablation results.

Methods	$Q_G\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$Q_S\uparrow$	$Q_{CB}\uparrow$	$VIF\uparrow$	$Q^{AB/F}\uparrow$	$MS2R\downarrow$
Full-DB	0.6276	0.9282	0.4897	0.8445	0.3943	0.3844	0.6991	0.9878
Full-SB	0.6260	0.9550	0.4857	**0.8680**	0.3996	0.3801	0.7108	0.9913
w/ IM	0.5880	0.7863	0.4690	0.8413	0.4017	0.3825	0.6889	0.9842
w/o SA	0.6080	0.8351	0.4861	0.8495	0.3984	0.3828	0.6975	0.9662
w/o MAFM	0.5972	0.7820	0.4676	0.8501	0.3962	0.3806	0.6876	0.9799
MAVFusion	**0.6325**	**0.9741**	**0.4947**	0.8504	**0.4036**	**0.3885**	**0.7182**	**0.8301**

This evaluation aims to compare the detection accuracy of various fused results across different scenarios. As illustrated in Fig. 7, in low-illumination environments where visible information is nearly entirely lost, object detection becomes heavily reliant on the thermal saliency provided by the infrared modality. Competing methods such as SAGE, GIFNet, and RFFusion fail to maintain high target-to-background contrast due to the excessive integration of non-informative visible noise, which significantly diminishes target saliency. Notably, our minimal intervention strategy for background regions suppresses interference from degraded visible inputs. By balancing modality integration and noise filtering, the method achieves the highest detection confidence across all targets, showing its advantage for high-level vision tasks.

4.5 Ablation Studies

Analysis of the Region Separating Strategy To verify the necessity of separating dynamic and static regions, we design three variant models for comparison:

- **Full-DB (Full Dynamic Branch)**: We remove the mask guidance and input the whole image into the dynamic branch for strong interaction, to evaluate the effect of global strong interaction.
- **Full-SB (Full Static Branch)**: We remove the dynamic branch and use only the static branch to process the whole image, to test the impact of lacking dynamic enhancement.
- **w/ IM (Inverted Mask)**: We reverse the mask logic, applying strong interaction to static regions and weak processing to dynamic regions, to verify our spatial feature allocation.

Analysis of Full-DB. As shown in Tab. 2, applying strong interaction to the whole image (Full-DB) increases computation cost, but does not improve performance. Salient and background regions compete during attention computation, which scatters the attention distribution and weakens local structures. As a result, the scores of Q_G, Q_M, and Q_S decrease.

Analysis of Full-SB. Removing the dynamic branch improves efficiency, but it ignores frame displacement in video sequences, as reflected by the clear deterioration of MS2R in Tab. 2. The model becomes a static weighting scheme, which causes overlap and conflicts around moving object boundaries. Although the Q_S score increases due to more stable backgrounds, this comes at the cost of dynamic object saliency.

Analysis of w/ IM. In this variant, static background regions, which only need simple information complement, are processed with strong interaction. In this case, Q_M, Q_P, and $Q^{AB/F}$ drop significantly. This shows that the strategy not only wastes computation, but also lets smooth infrared background information interfere with clear visible textures, causing blur and loss of contrast.

Analysis of the Sparse Attention We replace the original sparse attention with dense attention based on fully connected Softmax within each patch (w/o SA). As shown in Tab. 2, although the metric scores do not drop significantly, the results show degradation in details and contrast. This indicates that dense interaction covers more pixels, but processing redundant information does not effectively improve fusion performance and may smooth some local features.

Analysis of the MAFM Module To verify the effectiveness of the MAFM module, we remove it and replace it with the multi-frame alignment strategy used in UniVF (w/o MAFM). As shown in Tab. 2, most quality metrics decrease, while MS2R becomes worse. Without anchor features as alignment references, the ability of the model to integrate cross-modal features is weakened. In addition, removing the dynamic-static separation inside the module allows background features to interfere with dynamic regions.

4.6 Computational Efficiency Analysis

To evaluate efficiency, we compare MAVFusion with representative methods in terms of FLOPs, parameters, and FPS. RFFusion, a diffusion-based method, is not included due to its fixed-size input requirement and high computational cost.

As shown in Tab. 3, MAVFusion is much more efficient than video-fusion baselines. At 640×480, MAVFusion uses only 123.37G FLOPs, about 6.6% of VideoFusion (1874.00G) and 5.7% of UniVF (2164.07G), while achieving the highest FPS (14.16). At 1280×720, MAVFusion requires only 267.88G FLOPs, about 4.8% of VideoFusion and 3.8% of UniVF, and runs 2.22× faster than VideoFusion and 4.56× faster than UniVF. Moreover, when the number of pixels

Table 3: Comparison of computational complexity (FLOPs and parameters) and inference time (FPS) under different input resolutions.

Type	Methods	Pub	640×480			1280×720		
			FLOPs (G)	Params (M)	FPS	FLOPs (G)	Params (M)	FPS
Image	UP-Fusion	*AAAI 26*	953.86	154.82	1.80	2830.84	154.82	0.62
	TDFusion	*CVPR 25*	18.21	0.06	27.61	54.63	0.06	9.35
	SAGE	*CVPR 25*	20.34	0.136	153.45	61.00	0.136	51.63
	GIFNet	*CVPR 25*	226.12	0.82	5.68	678.32	0.82	2.64
	FreeFusion	*TPAMI 25*	451.65	5.67	12.81	1354.83	5.67	4.48
Video	VideoFusion	*CVPR 26*	1874.00	6.73	7.83	5623.00	6.73	2.59
	UniVF	*NeurIPS 25*	2164.07	9.20	4.01	7055.34	9.20	1.26
	MAVFusion	—	123.37	9.90	14.16	267.88	9.90	5.74

increases by 3×, the FLOPs of MAVFusion increase by only 2.17×, showing better scalability than VideoFusion and UniVF. Although its speed is lower than very lightweight image-fusion models such as TDFusion and SAGE, this is reasonable because video fusion requires temporal modeling and optical-flow estimation. Overall, MAVFusion achieves a favorable balance between fusion performance and computational efficiency.

4.7 Limitation

Although MAVFusion shows clear advantages in efficiency and performance, it still has some limitations. First, severe sensor noise or extreme degradation may still impair the extraction of reliable modality-specific features, making it difficult for the model to completely suppress degradation artifacts in highly corrupted regions. Second, while the region separation strategy greatly reduces fusion network computation, the extra cost of the front-end optical flow estimator remains a challenge. We plan to address both issues in future work.

5 Conclusion

This paper proposes MAVFusion, an efficient infrared and visible video fusion method based on motion-aware sparse interaction. It achieves high-fidelity fusion and frame-to-frame interaction while maintaining efficiency. To address cross-modal spatial misalignment and ghosting or blur caused by object motion, we design a lightweight Motion-Aware Feature Alignment Module. In the core interaction stage, video sequences are separated into dynamic and static regions: dynamic regions use mask-guided sparse attention for strong interaction to capture salient object features, while static regions use a lightweight weak interaction branch to reduce redundant computation and prevent semantic interference from non-informative modalities, maintaining background stability. Experiments on three benchmark datasets show that MAVFusion outperforms current state-of-the-art methods in both single-frame metrics and temporal consistency, achieving an effective balance between inference speed and fusion quality.

Acknowledgement

This work was supported in part by the Basic and Applied Basic Research of Guangdong Province under Grant 2023A1515140077, in part by the Natural Science Foundation of Guangdong Province under Grant 2024A1515011880, in part by the National Natural Science Foundation of China under Grant 52374166 and 62201149, in part by the Research Fund of Guangdong-Hong Kong-Macao Joint Laboratory for Intelligent MicroNano Optoelectronic Technology under Grant 2020B1212030010, and in part by the Yunnan Fundamental Research Projects under Grant 202301AV070004 and Grant 202501AS070123.

References

1. Bai, H., Zhang, J., Zhao, Z., Wu, Y., Deng, L., Cui, Y., Feng, T., Xu, S.: Task-driven image fusion with learnable fusion loss. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 7457–7468 (2025)
2. Cao, Z.H., Liang, Y.J., Deng, L.J., Vivone, G.: An efficient image fusion network exploiting unifying language and mask guidance. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
3. Chang, Z., Feng, Z., Yang, S., Gao, Q.: Aft: Adaptive fusion transformer for visible and infrared images. IEEE Trans. Image Process. **32**, 2077–2092 (2023)
4. Chen, J., Yu, W., Tian, X., Huang, J., Ma, J.: Mdbfusion: A visible and infrared image fusion framework capable for motion deblurring. In: IEEE Int. Conf. Image Process. pp. 1019–1025. IEEE (2024)
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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 28102–28112 (2025)
6. Ding, J., Zhang, H., Wang, Z., Xiao, J., Tian, X., Han, Z., Ma, J.: Cmvf: Cross-modal unregistered video fusion via spatio-temporal consistency. Inf. Fusion p. 104212 (2026)
7. Guo, X., Yang, F., Ji, L.: Mlf: A mimic layered fusion method for infrared and visible video. Infrared Phys. Technol. **126**, 104349 (2022)
8. Huang, J., Li, X., Tan, H., Yang, L., Wang, G., Yi, P.: Dednet: Infrared and visible image fusion with noise removal by decomposition-driven network. Measurement p. 115092 (2024)
9. Huang, J., Yan, P., Liu, J., Wu, J., Wang, Z., Wang, Y., Lin, L., Li, G.: Dreamfuse: Adaptive image fusion with diffusion transformer. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 17292–17301 (2025)
10. Li, H., Yang, Z., Zhang, Y., Jia, W., Yu, Z., Liu, Y.: Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. IEEE Trans. Pattern Anal. Mach. Intell. **47**(5), 3673–3690 (2025)
11. Li, H., Ma, H., Cheng, C., Shen, Z., Song, X., Wu, X.J.: Conti-fuse: A novel continuous decomposition-based fusion framework for infrared and visible images. Inf. Fusion **117**, 102839 (2025)
12. Li, H., Wu, X.J.: Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. Inf. Fusion **103**, 102147 (2024)

13. 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 Trans. Pattern Anal. Mach. Intell. (2023)
14. Li, W., Li, B., Song, H., Wang, P., Wang, Z.: Infrared and visible image fusion via iterative feature decomposition and deep balanced fusion. PR p. 113022 (2025)
15. Li, X., Li, X., Jiang, W.: Text-guided channel perturbation and pre-trained knowledge integration for unified multi-modality image fusion. In: AAAI. vol. 40, pp. 6521–6529 (2026)
16. Li, X., Li, X., Tan, T., Li, H., Ye, T.: Umcfuse: A unified multiple complex scenes infrared and visible image fusion framework. IEEE Trans. Image Process. **34**, 6231–6245 (2025)
17. Li, X., Li, X., Ye, T., Cheng, X., Liu, W., Tan, H.: Bridging the gap between multi-focus and multi-modal: a focused integration framework for multi-modal image fusion. In: IEEE Winter Conf. Appl. Comp. Vis. pp. 1628–1637 (2024)
18. Li, X., Liu, H., Li, X., Ye, T., Kuang, Z., Li, H.: Awm-fuse: multi-modality image fusion for adverse weather via global and local text perception. IEEE Trans. Image Process. **35**, 5151–5164 (2026)
19. Li, X., Liu, W., Li, X., Zhou, F., Li, H., Nie, F.: All-weather multi-modality image fusion: Unified framework and 100k benchmark. Inf. Fusion p. 104130 (2026)
20. Li, Z., Zhu, Z.L., Han, L.H., Hou, Q., Guo, C.L., Cheng, M.M.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 9801–9810 (2023)
21. Liu, C., Zhang, G., Zhao, R., Wang, L.: Sparse global matching for video frame interpolation with large motion. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 19125–19134 (2024)
22. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5802–5811 (2022)
23. 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 J. Autom. Sinica. (2024)
24. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2226–2235 (2025)
25. Liu, Z., Blasch, E., Xue, Z., Zhao, J., Laganiere, R., Wu, W.: Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 94–109 (2011)
26. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: Yolo26: key architectural enhancements and performance benchmarking for real-time object detection. arXiv preprint arXiv:2509.25164 (2025)
27. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 19559–19569 (2026)
28. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: Eur. Conf. Comput. Vis. pp. 36–54. Springer (2024)
29. Wang, Y., Zhuang, R., Zheng, H., He, X., Cao, K., Tu, X., Ding, X.: Self-supervised multiplex consensus mamba for general image fusion. arXiv preprint arXiv:2512.20921 (2025)

30. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 12637–12647 (2025)
31. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
32. Wang, Z., Zhang, J., Guan, T., Zhou, Y., Li, X., Dong, M., Liu, J.: Efficient rectified flow for image fusion. In: Adv. Neural Inform. Process. Syst. vol. 38, pp. 22927–22948 (2025)
33. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 17882–17891 (2025)
34. Xian, Y., Xu, Y., He, Y., Yang, Y.: From 2d grids to 1d tokens: Reforming shared representations for multimodal image fusion. arXiv preprint arXiv:2606.12303 (2026)
35. Xie, H., Sang, M., Zhang, Y., Yang, Y., Zhao, S., Zhong, J.: Rcvs: A unified registration and fusion framework for video streams. IEEE Trans. Multimedia **26**, 11031–11043 (2024)
36. Yang, B., Jiang, Z., Pan, D., Yu, H., Gui, G., Gui, W.: Lfdt-fusion: A latent feature-guided diffusion transformer model for general image fusion. Inf. Fusion **113**, 102639 (2025)
37. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 27026–27035 (2024)
38. Zhang, H., Cao, L., Zuo, X., Shao, Z., Ma, J.: Omnifuse: Composite degradation-robust image fusion with language-driven semantics. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
39. Zhang, X., Demiris, Y.: Visible and infrared image fusion using deep learning. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
40. Zhao, W., Cui, H., Wang, H., He, Y., Lu, H.: Freefusion: Infrared and visible image fusion via cross reconstruction learning. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
41. 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. In: Adv. Neural Inform. Process. Syst. vol. 38, pp. 128504–128535 (2025)
42. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5906–5916 (2023)
43. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 25912–25921 (2024)
44. 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: Int. Conf. Comput. Vis. pp. 8082–8093 (2023)
45. Zhu, Y., Wang, Q., Li, C., Tang, J., Gu, C., Huang, Z.: Visible–thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. PR **161**, 111330 (2025)