This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Streaming Diffusion Model for Fast Infrared and Visible Video Fusion

Jinyuan Liu[1], Ludan Sun[1], Tengyu Ma[1], Chunyan Yang[1],
Zhiying Jiang[2], Long Ma[1], Risheng Liu[1], Xin Fan[1]*

[1] School of Software Technology & DUT-RU International School of ISE, Dalian University of Technology

[2] College of Information Science and Technology, Dalian Maritime University

atlantis918@hotmail.com xin.fan@dlut.edu.cn

Abstract

Infrared and visible video fusion is pivotal for robust perceptual systems, aiming to synthesize a comprehensive video stream that leverages both thermal resilience and textured details. However, prevailing methods, by treating videos as sequences of independent frames, inherently introduce temporal incoherence, such as flickering and ghosting artifacts. While diffusion models possess strong generative priors to remedy this, their iterative nature is prohibitively slow for video. To resolve this fundamental dilemma, we propose a streaming diffusion model for efficient infrared and visible video fusion, termed SDMFusion. Our key insight is to exploit the generative prior of a pre-trained diffusion model into a one-step sampling framework, while explicitly modeling temporal dynamics. We design a memory-augmented latent pipeline where a temporal aggregation adapter aligns and propagates cross-frame features to ensure coherence, supported by a dedicated temporal consistency loss. This approach effectively decouples the challenge of achieving high fidelity from maintaining temporal stability. Extensive experiments on four benchmarks demonstrate that our method establishes a new state-of-the-art, generating fused videos with exceptional spatio-temporal consistency at a speed suitable for real-time application. The code is available at https://github.com/DandanYoung/SDMFusion.

1. Introduction

The pursuit of robust perceptual systems capable of operating under all environmental conditions has positioned infrared and visible video fusion as a cornerstone technology [1]. This process aims to synthesize a composite video stream that transcends the limitations of any single one [2, 3]. The thermal radiation information from infrared data, which is invariant to lighting variations and poor visibility, with the rich textual details and high spatial resolution inherent to the visible spectrum [4, 5]. The resulting fused video provides a more comprehensive and reliable representation of the scene, which is indispensable for a wide array of critical applications, including 24/7 surveillance, all-weather autonomous driving, and nighttime reconnaissance, to name a few [6, 7].

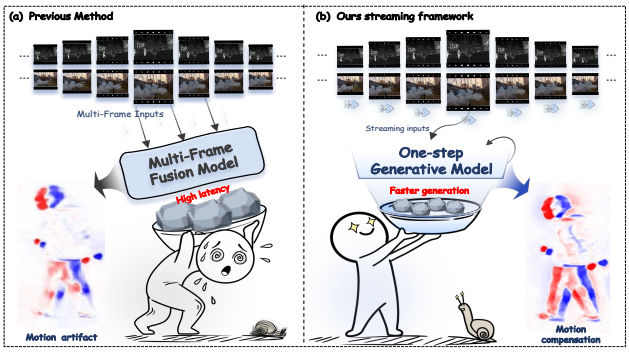


Figure 1. Left: Challenges of infrared and visible video fusion: afflicted with motion artifacts and burden of high latency. Right: Our method, SDMFusion, produces consistent backgrounds and coherent object motion with high efficiency.

Driven by these compelling needs, the field has witnessed significant advancements, which can be broadly categorized into two waves of development. The first and more mature wave focuses on ***image-level fusion***, where a plethora of deep learning-based methods have been proposed to generate a single, high-quality fused image from a pair of infrared and visible images [8, 9, 10]. ❶ However, these approaches predominantly treat video sequences as collections of independent frames, applying fusion in a simplistic frame-by-frame manner. This paradigm critically **overlooks the temporal dynamics** and **inter-frame dependencies inherent** in video data. Consequently, the fused video sequences often exhibit visually

*Corresponding author.

disturbing temporal flickering and object drift, which may introduce instability for downstream video tasks like tracking and behavior recognition.

Acknowledging this gap, a second, nascent wave of research has begun to explore ***video fusion*** techniques [11, 12]. These pioneering works represent a step in the right direction by attempting to incorporate temporal information. (2) Yet, they often rely on simplistic mechanisms such as temporal averaging or basic recurrent units, which are insufficient for capturing the complex, non-linear motion patterns and long-range dependencies present in real-world videos. Frequently, these methods result in over-smoothed outputs, blurred motion trails, or residual artifacts around moving objects, achieving **limited success in ensuring genuine temporal coherence**.

Concurrently, the field of generative models has been revolutionized by the advent of Diffusion Models (DMs), which have set new standards for image generation fidelity due to their powerful ability to model complex data distributions [13, 14]. This presents a tantalizing opportunity for the fusion community. (3) However, the iterative denoising process that underpins DMs incurs substantial computational overhead, rendering them **prohibitively slow for processing video sequences**, which comprise dozens or hundreds of frames. While recent one-step distillation techniques have emerged to dramatically accelerate sampling for images, they are inherently temporally-naïve [15]. Therefore, a direct application would merely produce a sequence of high-quality but temporally inconsistent fused images, failing to solve the core challenge of video fusion.

To meet real-time requirements while fully exploiting the temporal coherence of video sequences, we **propose a streaming diffusion model** for fast infrared–visible video fusion. Specifically, a dual-branch encoder first extracts cross-modal features and compresses them into a latent space to **curb computational burden**. To capitalize on generative priors, a streaming diffusion model (SDM) with one-step sampling is introduced, which requires the current frame and a compact memory unit, thereby **reducing latency**. In the SDM, the denoising U-Net is frozen during training to preserve its generative capability learned from large-scale data, and a temporal aggregation adapter is inserted after the U-Net decoder layers to **aggregate cross-frame features and update the memory**. The refined features are then restored to the original resolution and reconstructed into fused frames by the decoder. In addition, we introduce a temporal consistency loss that **constrains the stability of temporal feature propagation** during training. Extensive experiments demonstrate the effectiveness of SDMFusion and its promise for real-world deployment. The contributions of this work are summarized below.

- We propose a streaming diffusion model that captures dynamic temporal dependencies across video sequences, explicitly models cross-frame relations, and generates temporally coherent videos.
- We align adjacent-frame features via optical flow and aggregate them along the temporal axis, while a temporal-consistency loss reduces flicker and drift to improve cross-frame coherence.
- We design a cross-frame feature aggregation adapter within the one-step diffusion model to refine temporal features and reduce motion-induced artifacts.
- Extensive experiments on multiple datasets demonstrate that the method stably produces fused videos with consistent backgrounds and coherent object motion.

2. Related work

2.1. Frame-by-frame image fusion

With the rapid advance of deep learning, infrared and visible image fusion methods built on continuously evolving neural network architectures also achieve substantial progress [16, 17]. Among them, autoencoder-based approaches [18, 19] employ pre-trained encoders to map multimodal inputs to latent codes and merge these representations through hand-crafted fusion rules. To overcome the limited representational capacity of autoencoders, convolutional neural networks [5, 20, 21] and transformers [22, 23] are adopted due to their strong local and global modeling abilities, which improve fusion performance but depend heavily on loss design. Subsequently, generative adversarial networks [24, 25] learn the data distribution of the fusion results through adversarial optimization between a generator and a discriminator, but the gradients provided by the discriminator can be unstable and prone to degradation, reducing training controllability. In contrast, diffusion models approximate the data distribution through conditional denoising in a step-by-step manner and provide dense and stable training signals [26, 27], but the absence of reference images limits their application in IVIF. Regarding this issue, Zhao *et al.* [10] formulate multi-modal fusion as conditional generation under a DDPM framework, leveraging its generative prior without requiring training, at the cost of high inference overhead due to the multi-step denoising process. Yi *et al.* [28] integrate the results of existing fusion methods as prior knowledge to generate pseudo-labels for training, which avoids the instability of adversarial learning but limits performance due to reliance on the prior.

2.2. Video restoration

Although frame-by-frame methods perform well on static frames, they ignore crucial temporal dependencies across video sequences [11]. In adjacent frames, the background is highly similar, and objects exhibit continuous motion in a specific direction [29]. Hence, fully exploiting inter-frame

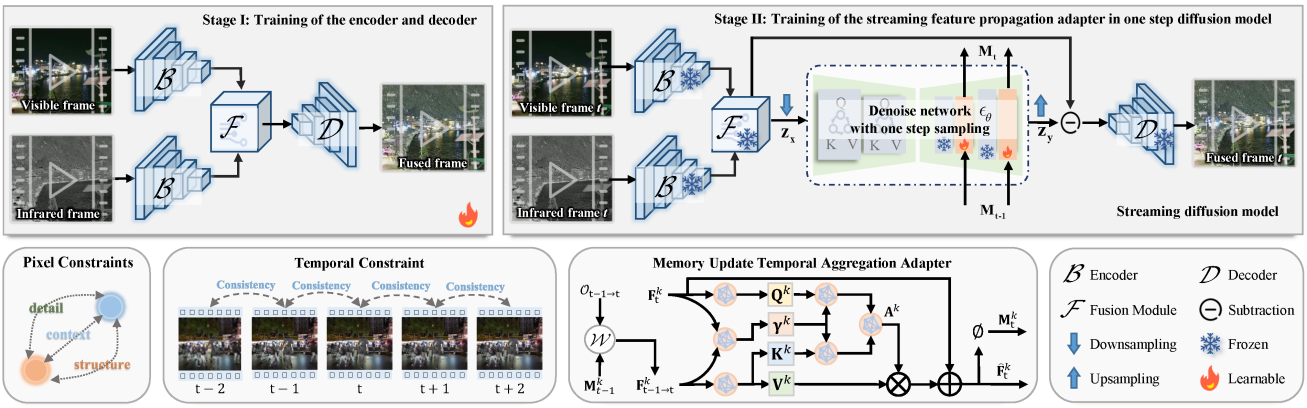


Figure 2. Overall pipeline of the proposed SDMFusion. Stage I trains the image-level fusion backbone, while Stage II introduces the streaming diffusion model with memory update temporal aggregation adapter to produce temporally coherent fused videos efficiently.

features is crucial to compensate for occlusion and restore degraded information [30]. By leveraging feature similarity between adjacent frames, MIA-VSR [31] selects enhanced features from the previous frame for interaction with the current frame, thereby propagating effective information over time. STCNet [32] introduces event streams as temporal motion cues and achieves video deblurring through cross-modal spatial fusion and temporal collaboration. Recent studies, such as RCVS [11], show that temporal cues are crucial in infrared–visible scenarios and mitigate flicker artifacts caused by per-frame processing. TemCoCo [12] integrates visual–semantic collaboration with explicit temporal modeling within a video fusion network, achieving temporal coherence in the fused videos. UniVF [1] performs explicit temporal modeling by aggregating information from adjacent frames through multi-frame learning and optical flow-based feature warping, but repeated processing of multiple frames leads to low feature utilization and increased computational cost. To this end, we propose a streaming framework based on a one-step diffusion model that directly produces the current fused frame during inference without requiring the entire sequence as input.

3. Motivation

Static frame-based fusion frameworks have been extensively studied [33, 34], but their direct application to video fusion faces several bottlenecks. First, processing frames independently ignores cross-temporal dependencies and fails to address flicker and artifacts caused by object motion [11]. Multi-frame input schemes can compensate information via feature reuse, but naive temporal stacking increases computational cost during training and latency at inference [1, 12]. Meanwhile, diffusion models have demonstrated strong generative priors for detail restoration, yet standard multi-step diffusion remains computationally expensive, unstable when directly adapted to video fusion, and lacks designs tailored to cross-modal constraints and temporal structure [10, 13]. To address these issues, we propose a streaming diffusion-based video fusion framework that leverages generative priors and inter-frame feature propagation, supervised by a temporal consistency constraint, to mitigate flicker and motion-induced artifacts, while a one-step sampling strategy substantially reduces the inherent computational overhead of diffusion models.

4. Methodology

This section elaborates how the proposed SDMFusion models inter-frame temporal correlations to compensate for information loss and artifacts caused by object motion, and employs a streaming framework to generate coherent videos in a fast and stable manner. Given an aligned infrared and visible video pair $\{(\mathbf{I}_t, \mathbf{V}_t)\}_{t=1}^{T}$, $\mathbf{I}_t, \mathbf{V}_t \in \mathbb{R}^{H\times W}$, our goal is to generate a fused sequence $\{\hat{\mathbf{Y}}_t\}_{t=1}^{T}$, $\hat{\mathbf{Y}}_t \in \mathbb{R}^{H\times W}$. Instead of directly training a heavy video diffusion model, we adopt a two-stage design. Stage I learns a cross-modal encoder–decoder on static infrared and visible image pairs. Stage II freezes this backbone and introduces a streaming diffusion model with flow-guided feature propagation to refine latent representations over time. This decoupling stabilizes training, exploits strong image-level priors, and yields efficient and fast video fusion. Our framework is inspired by recent one-step latent diffusion paradigms that distill multi-step diffusion into a single residual prediction while preserving generative priors.

4.1. Preliminaries

Stable Diffusion (SD) operates in a VAE $\mathcal{V}$ latent space and learns to invert a fixed Gaussian noising process, which provides strong priors for preserving global fidelity and restoring fine textures. Let $\mathbf{z}_0 \in \mathbb{R}^{C\times H\times W}$ denote a clean latent. The forward diffusion corrupts $\mathbf{z_0}$ as $q(\mathbf{z_t} \mid \mathbf{z_0}) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}\big)$ with $\bar{\alpha}_z = \prod_{s=1}^{z} \alpha_s$. In the reverse

process, the denoiser ϵ_θ yields the standard reverse mean update through the following formula:

$$\mu_\theta(\mathbf{z}_\mathtt{t}, \mathtt{t}, c) = \frac{1}{\sqrt{\alpha_\mathtt{t}}}\Big(\mathbf{z}_\mathtt{t} - \tfrac{1-\alpha_\mathtt{t}}{\sqrt{1-\bar{\alpha}_\mathtt{t}}}\,\epsilon_\theta(\mathbf{z}_\mathtt{t}, \mathtt{t}, \mathtt{c})\Big), \quad (1)$$

and the reverse transition is modeled as:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_\mathtt{t}, \mathbf{c}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_\mathtt{t}, \mathtt{t}, \mathbf{c}),\, \sigma_\mathtt{t}^2\mathbf{I}\big), \quad (2)$$

where $\sigma_\mathtt{t}$ is determined by the chosen sampler. Conventional diffusion samplers such as DDPM and DDIM, generate a sample by iterating over T steps from $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$:

$$\mathbf{z}_{t-1} \leftarrow \mathcal{S}\big(\mathbf{z}_\mathtt{t},\, \mathtt{t},\, \epsilon_\theta(\mathbf{z}_\mathtt{t}, \mathtt{t}, \mathbf{c})\big), \qquad \mathtt{t} = T, \ldots, 1, \quad (3)$$

where $\mathcal{S}$ denotes one reverse-update step. Although such multi-step sampling attains excellent fidelity and detail, multi-step sampling incurs substantial inference latency and limits real-time video fusion.

Latent diffusion pipelines often rely on a VAE trained on a single modality distribution such as the visible domain $\mathcal{P}_V$. The visible-only VAE can introduce distribution shift when directly applied to fused content with distribution $\mathcal{P}_F$ stemming from both infrared and visible cues. $\mathcal{KL}\big(p_F(\mathbf{z}_\mathtt{t}) \,\|\, p_V(\mathbf{z}_\mathtt{t})\big)$ is large and $\mathcal{V}^{-1} \circ \mathcal{V}$ is misaligned on fusion latents, manifesting as spatial misregistration and ghosting after decode. Empirically, we observe that the encoder trained solely on visible images tends to underrepresent thermal saliency, while the decoder amplifies visible textures in misaligned locations, producing spatial artifacts in the fused output. To resolve the mismatch, we replace the off-the-shelf VAE with a simple and task-adaptive encoder-decoder that follows a DCEvo-style design for domain adaptation. It is trained aligned with the distribution of fused latents rather than visible-only latents.

4.2. One step diffusion model

Given an aligned infrared and visible frame pairs $(\mathbf{I}_\mathtt{t}, \mathbf{V}_\mathtt{t})$ at time $\mathtt{t}$, we first extract cross-modal features with the Stage I backbone and fuse them in feature space $\mathbf{G}_\mathtt{t}$, where $\mathbf{G}_\mathtt{t} = \mathcal{F}(\mathcal{B}(\mathbf{I}_\mathtt{t}),\, \mathcal{B}(\mathbf{V}_\mathtt{t}))$. To reduce computation inside diffusion, we downsample the fused features by a fixed spatial factor $\mathtt{s}$ and $\mathbf{z}_\mathtt{t} = Downsampling(\mathbf{G}_\mathtt{t}) \in \mathbb{R}^{\mathtt{C}\times\frac{\mathtt{H}}{\mathtt{s}}\times\frac{\mathtt{W}}{\mathtt{s}}}$. We reuse a pretrained SD denoising U-Net ϵ_θ and insert adapter blocks $\{\mathcal{A}^{(k)}\}$ after its decoder layers to inject temporal perception aggregation while keeping the backbone fixed to retain the generative ability of the diffusion model. To overcome the latency bottleneck, we adopt a single-step residual correction in latent space using a fixed diffusion time. At a fixed diffusion time $\hat{\mathtt{t}}$, the denoiser predicts a residual on the downsampled latent feature by $\mathbf{z}_\mathtt{x} = \epsilon_\theta(\mathbf{z}_\mathtt{t};\, \hat{\mathtt{t}},\, \{\mathcal{A}^{(k)}\}\Big)$, and we perform one step latent update using the standard closed-form relation:

$$\mathbf{z}_\mathtt{y} = \frac{\mathbf{z}_\mathtt{x} - \sqrt{1-\bar{\alpha}_{\hat{\mathtt{t}}}}\;\mathbf{r}_\mathtt{t}}{\sqrt{\bar{\alpha}_{\hat{\mathtt{t}}}}}. \quad (4)$$

Where $\bar{\alpha}_{\hat{\mathtt{t}}} = \prod_{\mathtt{s}=1}^{\hat{\mathtt{t}}} \alpha_\mathtt{s}$ follows the same schedule as in the preliminaries, $\mathtt{r}_\mathtt{t} = \epsilon_\theta\big(\mathbf{Z}_\mathtt{x};\, \mathtt{c},\, \hat{\mathtt{t}}\big)$ denotes the residual predicted by the denoiser. Eq. (4) replaces the entire iterative sampling chain, reducing inference to a single forward pass while retaining the generative priors responsible for fidelity and detail. Subsequently, the refined latent is brought back to the original feature scale and merged with a residual skip to preserve low-frequency structure in a stable reconstruction process. Finally, the image decoder $\mathcal{D}$ reconstructs the fused frame from latent space.

4.3. Streaming diffusion model

Despite promising results from several video-fusion methods, multi-frame inputs substantially increase the computational burden during both training and inference due to repeated backbone evaluations and additional memory traffic. The strict dependence on adjacent frames also introduces pipeline latency, which undermines real-time deployment. We therefore propose a streaming framework that only requires the current frame, previous frame, and a prior memory. In practice, an optical-flow estimator predicts the flow map $\mathcal{O}_{\mathtt{t}-1\to\mathtt{t}}$. The memory prior updated from the previous step, $\mathbf{M}_{\mathtt{t}-1}$, is aligned to time $\mathtt{t}$ via warping as $\tilde{\mathbf{F}}_{\mathtt{t}-1\to\mathtt{t}} = \mathcal{W}(\mathbf{M}_{\mathtt{t}-1},\, \mathcal{O}_{\mathtt{t}-1\to\mathtt{t}})$, and injected into the denoising U-Net through adapter layers placed after decoder blocks to aggregate temporal information.

Let $\mathbf{F}_\mathtt{t}^{(k)} \in \mathbb{R}^{C_k\times H_k\times W_k}$ denote the current decoder feature at scale k, and $\tilde{\mathbf{F}}_{\mathtt{t}-1\to\mathtt{t}}^{(k)}$ the flow-aligned memory at the same scale. Each adapter performs content-aware aggregation through a lightweight cross-attention followed by a gated residual fusion. Specifically, the query $\mathbf{Q}^{(k)}$ is projected from the current feature $\mathbf{F}_\mathtt{t}^{(k)}$, while the key $\mathbf{K}^{(k)}$ and value $\mathbf{V}^{(k)}$ are derived from the flow-aligned memory $\tilde{\mathbf{F}}_{\mathtt{t}-1\to\mathtt{t}}^{(k)}$ via 1×1 convolutions $\phi_q^{(k)}$, $\phi_k^{(k)}$, and $\phi_v^{(k)}$, respectively. To modulate the temporal interaction, a learned gate $\boldsymbol{\gamma}^{(k)}$ is computed by:

$$\boldsymbol{\gamma}^{(k)} = \sigma(\phi_g^{(k)}(\mathrm{Cat}(\mathbf{F}_\mathtt{t}^{(k)}, \tilde{\mathbf{F}}_{\mathtt{t}-1\to\mathtt{t}}^{(k)}))), \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function and Cat denotes concatenation. The gated attention and feature aggregation process is then formulated as:

$$\mathbf{A}^{(k)} = \mathrm{Softmax}\Big(\frac{(\mathbf{Q}^{(k)} \odot \boldsymbol{\gamma}^{(k)})(\mathbf{K}^{(k)} \odot \boldsymbol{\gamma}^{(k)})^\top}{\sqrt{C}}\Big), \quad (6)$$

$$\hat{\mathbf{F}}_\mathtt{t}^{(k)} = \mathbf{F}_\mathtt{t}^{(k)} + \mathbf{A}^{(k)}\mathbf{V}^{(k)}, \quad (7)$$

where C is the channel dimension and $\odot$ is element-wise multiplication. This design lets the current feature query motion-aligned memory and selectively absorb complementary cues with the learned gate, which mitigates motion-induced artifacts and suppresses temporal flicker. Finally, the adapter updates the memory state as $\mathbf{M}_\mathtt{t}^{(k)} =$

Table 1. Quantitative comparison with state-of-the-art fusion approaches on the HDO, VTMOT, NOT-156, and M3SVD datasets. Red highlights the best performance, while green denotes the second-best results.

Method	Pub.	HDO			M3SVD			VTMOT			NOT-156		
		SCD↑	VIF↑	mSSIM↑	SCD↑	VIF↑	mSSIM↑	SCD↑	VIF↑	mSSIM↑	SCD↑	VIF↑	mSSIM↑
CDDFuse [22]	*CVPR'23*	1.509	0.715	1.134	1.621	0.866	0.925	1.585	0.907	0.983	1.614	0.890	1.091
LRRNet [35]	*TPAMI'23*	1.153	0.528	0.980	1.314	0.669	0.802	1.040	0.628	0.827	1.385	0.575	0.977
Mask-Dif [27]	*TPAMI'24*	1.585	0.680	1.061	1.688	0.769	0.815	1.600	0.969	0.933	1.561	0.796	1.032
TIMFus [36]	*TPAMI'24*	1.146	0.633	0.942	1.461	0.659	0.900	0.563	0.683	0.753	1.177	0.754	0.961
DCEvo [37]	*CVPR'25*	1.604	0.765	1.127	1.747	0.871	0.933	1.550	1.026	1.032	1.585	0.913	1.087
GIFNet [38]	*CVPR'25*	1.610	0.573	1.087	1.546	0.582	0.894	1.545	0.764	0.942	1.607	0.669	1.018
SAGE [39]	*CVPR'25*	1.655	0.599	1.065	1.760	0.771	0.932	1.556	0.750	0.952	1.611	0.658	1.010
UniVF [1]	*NeurIPS'25*	1.496	0.790	1.113	1.721	0.914	0.939	1.558	1.068	1.037	1.509	0.912	1.075
TemCoCo [12]	*ICCV'25*	0.618	0.632	0.953	1.553	0.678	0.894	1.307	0.716	0.915	0.616	0.455	0.795
Ours	—	1.572	0.794	1.135	1.776	0.926	0.955	1.583	1.026	1.059	1.519	0.937	1.094

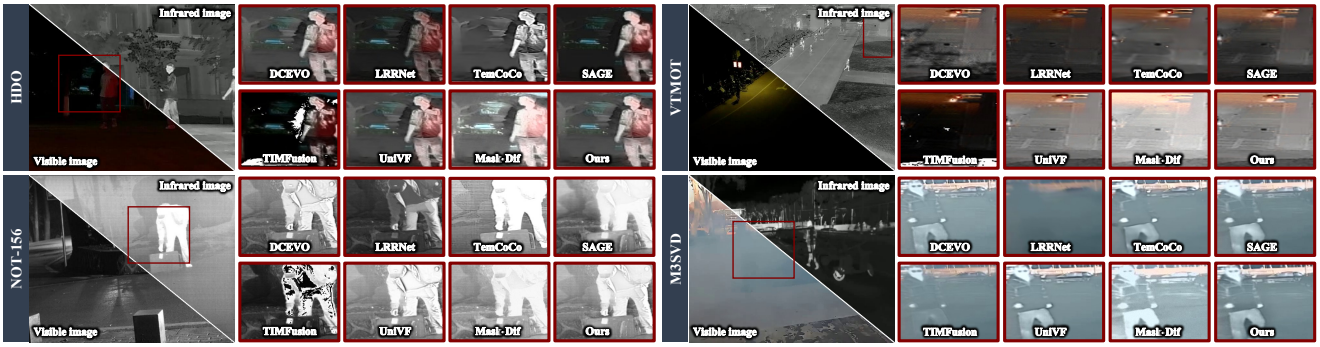


Figure 3. Qualitative comparisons between our method and existing fusion approaches on the four datasets.

$\phi_m^{(k)}(\hat{\mathbf{F}}_{\mathtt{t}}^{(k)})$, and the multi-scale set $\{\mathbf{M}_{\mathtt{t}}^{(k)}\}_k$ serves as the prior for the next frame $\mathtt{t}+1$. This streaming, memory-augmented scheme requires only the current frame and a compact prior, thereby reducing pipeline latency while maintaining high-fidelity temporal coherence.

4.4. Loss function

To train the one-step diffusion model, we employ a combination of modality reconstruction, structure-preserving fusion, and temporal consistency losses. For each time step $\mathtt{t}$, given the visible frame $\mathbf{V}_{\mathtt{t}}$, infrared frame $\mathbf{I}_{\mathtt{t}}$ and the fused prediction $\hat{\mathbf{Y}}_{\mathtt{t}}$, we first impose two modality reconstruction losses that encourage the fused result to be consistent with inputs. We combine the MSE loss and SSIM loss terms:

$$\mathcal{L}_{\mathrm{V,I}} = L_{\mathrm{SSIM}}\left(\mathbf{Y}_{\mathtt{t}}, \hat{\mathbf{Y}}_{\mathtt{t}}\right) + \left\|\mathbf{Y}_{\mathtt{t}} - \hat{\mathbf{Y}}_{\mathtt{t}}\right\|_2^2, \tag{8}$$

where $L_{\mathrm{SSIM}}(\cdot,\cdot)$ denotes the structural similarity loss and $\|\cdot\|_2$ is the ℓ_2 norm, $\mathbf{Y}_{\mathtt{t}}$ denote the $\mathbf{I}_{\mathtt{t}}$ or $\mathbf{V}_{\mathtt{t}}$. To preserve both intensity and structures from source intputs, we adopt a fusion loss $\mathcal{L}_{\mathrm{fus}}$ built on luminance and gradient cues. Let $\mathbf{V}_{\mathtt{t}}^{Y}$ be the luminance channel of the visible image and $\nabla(\cdot)$ the Sobel-gradient operator. We form intensity and gradient references by taking element-wise maxima and define:

$$\mathcal{L}_{\mathrm{fus}} = \left\|\max\left(\mathbf{V}_{\mathtt{t}}^{Y}, \mathbf{I}_{\mathtt{t}}\right) - \hat{\mathbf{Y}}_{\mathtt{t}}\right\|_1 + \left\|\max\left(\nabla\mathbf{V}_{\mathtt{t}}^{Y}, \nabla\mathbf{I}_{\mathtt{t}}\right) - \nabla\hat{\mathbf{Y}}_{\mathtt{t}}\right\|_1, \tag{9}$$

which encourages the fused image to inherit the stronger intensity responses and sharper edges from either modality.

To further enhance temporal coherence, during the second stage of training, we introduce a temporal consistency loss $\mathcal{L}_{\mathrm{tc}}$ defined on the fused video $\{\hat{\mathbf{Y}}_{\mathtt{t}}\}_{\mathtt{t}=1}^{\mathtt{T}}$ and the estimated forward flows $\{\mathbf{O}_{\mathtt{t}-1\to\mathtt{t}}\}_{\mathtt{t}=2}^{\mathtt{T}}$ by pretrained SpyNet [40]. Let $\mathcal{W}(\cdot, \mathbf{O}_{\mathtt{t}-1\to\mathtt{t}})$ be the warping operator and $\mathbf{M}_{\mathtt{t}}$ a binary validity mask obtained by warping an all-one map. For T frames, we compute:

$$\mathcal{L}_{\mathrm{tc}} = \frac{1}{T-1}\sum_{\mathtt{t}=2}^{T} \frac{\left\|\left(\hat{\mathbf{Y}}_{\mathtt{t}} - \mathcal{W}(\hat{\mathbf{Y}}_{\mathtt{t}-1}, \mathbf{O}_{\mathtt{t}-1\to\mathtt{t}})\right)\odot\mathbf{M}_{\mathtt{t}}\right\|_1}{\sum\mathbf{M}_{\mathtt{t}} + \varepsilon}, \tag{10}$$

where $\odot$ denotes element-wise multiplication and ε is a small constant for numerical stability. This term encourages the current fused frame to be consistent with the motion-compensated previous frame in valid regions, thereby suppressing flicker and drift.

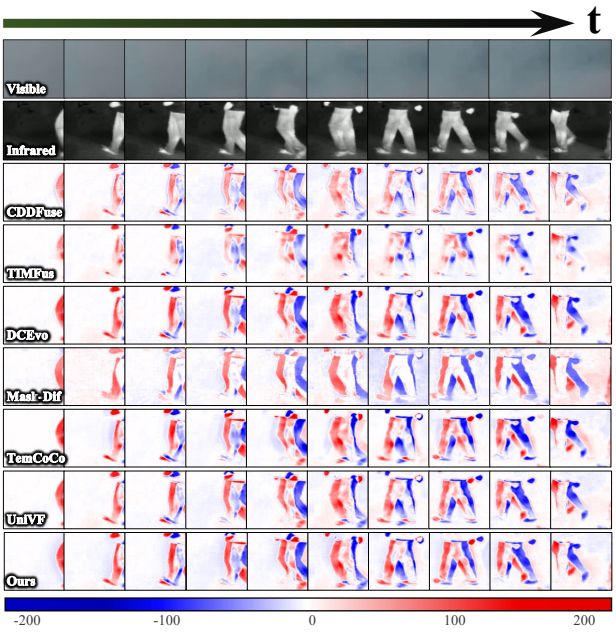


Figure 4. Frame-wise difference visualization on the M3SVD dataset for temporal consistency evaluation.

The overall training objective for SDMFusion is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{V,I}}\mathcal{L}_{\text{V,I}} + \lambda_{\text{fus}}\mathcal{L}_{\text{fus}} + \lambda_{\text{deco}}\mathcal{L}_{\text{deco}} + \lambda_{\text{tc}}\mathcal{L}_{\text{tc}}, \tag{11}$$

where $\lambda_{\text{V,I}}$, λ_{fus}, λ_{deco}, and λ_{tc} are trade-off weights, $\mathcal{L}_{\text{deco}}$ is proposed by [22]. This composite loss guides the model to simultaneously preserve modality-specific structures, produce visually sharp fused frames, and maintain temporal coherence across the video sequence.

5. Experiment

In this section, we evaluate the proposed method using extensive visual comparisons together with a comprehensive set of quantitative metrics. We first present the experimental settings and implementation details. Then we report fusion performance on four benchmark datasets to demonstrate that our approach consistently produces superior fused results. To further clarify the contribution of each component, we conduct ablation studies on the streaming diffusion module and the optical-flow-guided feature propagation. Finally, we include an efficiency analysis showing that the proposed framework achieves lower latency and is well suited for practical video fusion scenarios.

5.1. Experimental settings and details

We conduct experiments on four infrared and visible video fusion benchmarks, HDO [11], M3SVD [41], VTMOT [1] and NOT-156 [42]. These datasets cover diverse environments and motion patterns, including urban surveillance scenes, multi sensor driving scenarios and sequences with complex dynamic backgrounds. The batch size in Stage I is set to 12, while Stage II uses a batch size of 1 with a sequence length of 6. In both stages, the network parameters are optimized using AdamW with an initial learning rate of 1×10^{-4}, and a cosine annealing schedule that decays the learning rate to 1×10^{-5}. Our framework is implemented in PyTorch and runs on a workstation equipped with a single NVIDIA GeForce RTX 5090 GPU.

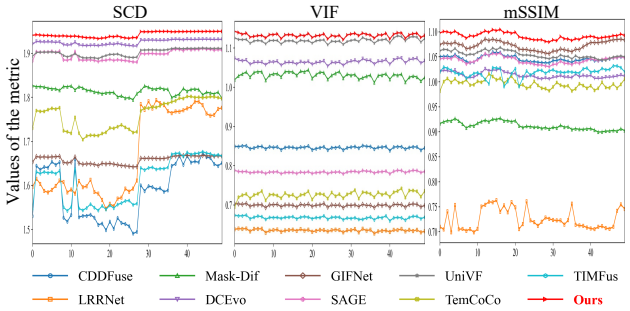


Figure 5. Visualization of frame-wise SCD, VIF, and mSSIM scores on a video sequence from the M3SVD dataset.

5.2. Results of multi-modality video fusion

In this section, we compare the performance of the proposed method with 7 state-of-the-art image fusion methods, including CDDFuse [22], LRRNet [35], Mask-Dif [27], TIMFus [36], DCEvo [37], GIFNet [38], SAGE [39], and two video fusion methods, UniVF [1] and TemCoCo [12].

Quantitative Comparisons. We analyze the performance of our proposed method based on the objective metrics SCD, VIF, and mSSIM. These metrics are essential for evaluating the quality of fused images, including their spatial consistency, visual quality, and structural similarity. As shown in Table 1, our method achieves the best results in most of the metrics, with only a slight difference in the remaining ones. In summary, the good performance on SCD, VIF, and mSSIM indicate that our method outperforms existing methods in terms of spatial consistency, visual fidelity, and structural accuracy.

Qualitative Comparisons. Additionally, we selected one fusion result from each of the four datasets for visualization analysis, as shown in Figure 3. Our method better preserves the thermal radiation information from the infrared input and the texture details from the visible image. The higher contrast achieved is also beneficial for downstream tasks.

5.3. Evaluation of temporal consistency

To examine the temporal consistency and continuity across video frames, we conduct a frame-difference analysis on a sequence from the M3SVD dataset. Specifically, for each

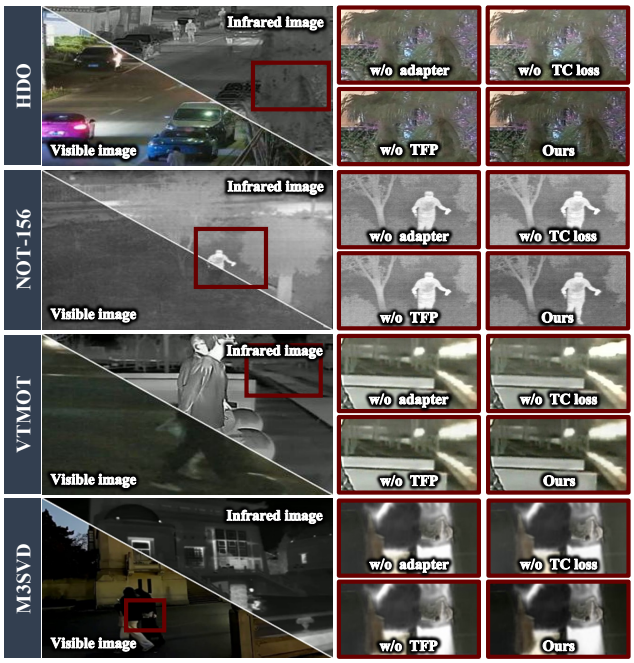


Figure 6. Qualitative comparisons of different ablative variants on the HDO, VTMOT, NOT-156, and M3SVD datasets.

pair of consecutive frames, we compute the pixel-wise intensity difference map by subtracting the previous frame from the subsequent one. We then apply a color-coded visualization to this difference map, where red denotes positive differences, blue denotes negative differences, and larger absolute values correspond to deeper colors. As illustrated in Figure 4, static regions such as background areas exhibit minimal temporal variation and therefore appear close to white. Compared with Mask-Difuser, CDDFuse, and TIM-Fusion, our method yields cleaner and more stable background regions, demonstrating stronger suppression of temporal noise. In addition, the motion boundaries of the moving object are more sharply captured by our method, reflecting superior sensitivity to meaningful temporal changes.

To further assess temporal stability over longer sequences, we also visualize the frame-wise trajectories of three widely used video-quality metrics, including SCD, VIF, and mSSIM, on consecutive frames of the M3SVD dataset. As shown in Figure 5, our method not only achieves the highest overall scores but also maintains consistently stable trends over time, indicating robust temporal coherence and reduced frame-to-frame fluctuations. These results highlight the strong capability of our proposed framework to preserve temporal consistency in video fusion tasks.

5.4. Ablation experiment

To analyze the contribution of each key component in SDMFusion, we conducted ablation studies on four datasets.

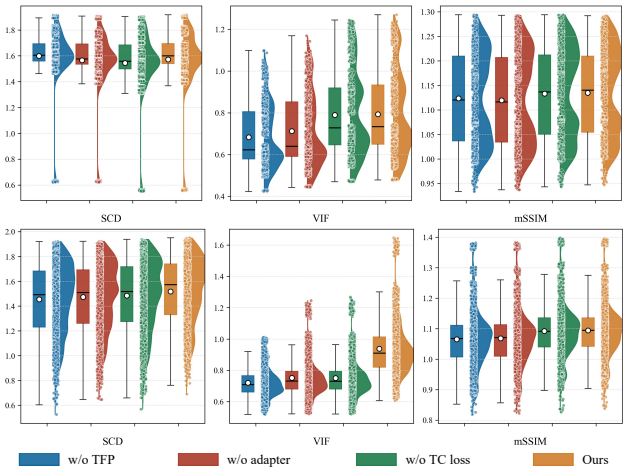


Figure 7. Violin plots of SCD, VIF, and mSSIM for different ablation variants on the HDO (top) and NOT-156 (bottom) datasets.

Table 2. Quantitative results of ablation study on the M3SVD and VTMOT datasets. **Bold** values indicate the best performance.

Method	M3SVD			VTMOT		
	SCD↑	VIF↑	mSSIM↑	SCD↑	VIF↑	mSSIM↑
w/o TFP	1.771	0.873	0.945	1.569	1.018	1.054
w/o adapter	1.743	0.897	0.946	1.562	1.024	1.057
w/o TC loss	1.769	0.920	0.950	1.576	1.018	1.057
Ours	**1.776**	**0.926**	**0.955**	**1.583**	**1.026**	**1.059**

Specifically, we removed the feature injection of the temporal prior $\mathbf{M}_t^{(k)}$ and replaced it with features $\mathbf{F}_t^{(k)}$ extracted from the current frame. We also removed the adapter after the UNet decoding layer and replaced it with a standard convolution. Finally, we removed the temporal consistency (TC) loss function to validate its effect on enforcing temporal consistency in the fused video. Table 2 reports the quantitative results on the M3SVD and VTMOT datasets. As can be observed, removing any single component leads to a noticeable performance drop. Figure 7 shows the distribution of metrics on the HDO and NOT-156 datasets, where the full SDMFusion achieves the best performance. These objective results demonstrate the effectiveness of each key part of the proposed SDMFusion. In addition, we select one fused frame from each of the four datasets for visual analysis. As shown in Figure 6, the complete SDMFusion produces clearer textures and higher contrast.

5.5. Object tracking

To evaluate the applicability on downstream tasks of fused video generated by our proposed method, we select several video sequences from the NOT-156 dataset to conduct object tracking. For a fair comparison, ByteTrack tracker



Figure 8. Qualitative tracking results on the NOT-156 dataset.

Table 3. Quantitative comparison on tracking task across different fusion methods on the NOT-156 dataset.

Method	AUC↑	SR@0.5↑	SR@0.75↑	DP@20↑
CDDFuse	0.3529	0.3911	0.1467	0.3556
LRRNet	0.3614	0.4000	0.1400	0.3500
Mask-Dif	0.3774	0.4277	0.1785	0.3836
TIMFus	0.3140	0.3618	0.1446	0.3104
DCEvo	0.3673	0.4200	0.1550	0.3800
GIFNet	0.3537	0.3936	0.1655	0.3500
SAGE	0.3491	0.3888	0.1506	0.3501
UniVF	0.3553	0.3826	0.1363	0.3607
TemCoCo	0.3102	0.3597	0.1499	0.3328
Ours	0.3799	0.4350	0.1700	0.3900

[43] with the YOLOv11n [44] pretrained detector is directly adopted to perform tracking without training. To conduct an objective evaluation, we introduce the standard metrics of area under the curve (AUC), success rate (SR), and distance precision (DP) [45] as our evaluation criteria. AUC represents the area under the success rate curve. SR indicates the proportion of successfully tracked frames where the overlap exceeds a specified threshold. We calculate SR using two overlap thresholds: 0.5 and a more stringent 0.75. DP@20 (distance precision at 20 pixels) measures localization accuracy by calculating the percentage of frames where the Euclidean distance between the predicted and ground-truth bounding box centers is less than 20 pixels.

As summarized in Table 3, our method achieves the best performance on AUC, SR@0.5 and DP@20 and second-best performance behind the Mask-Dif on SR@0.75, indicating superior tracking robustness, higher success rates, lower localization errors, and more accurate and tightly fitted bounding boxes. To further illustrate the effectiveness of proposed fusion framework, we visualize the tracking results on a fused frame as shown in Figure 8. Our method successfully identifies and maintains a confidence score of 0.90, which strongly benefits from the proposed fusion model's ability to recover fine-grained texture details and suppress noise effectively, thereby providing a more reliable input for the downstream tracker.

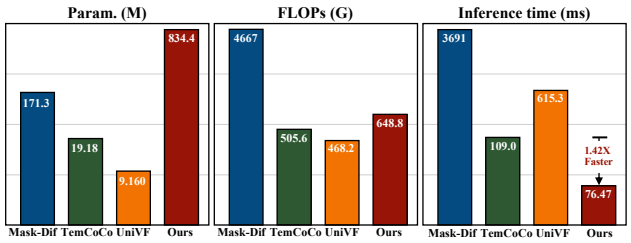


Figure 9. Efficiency comparison of representative fusion methods.

5.6. Efficiency analysis

To evaluate computational efficiency, we compare our method with two video fusion and one diffusion methods on the VTMOT dataset. As shown in Figure 9, our method achieves the fastest inference speed by a significant margin, whose total inference time is more than 1.42× faster than the second-best competitor. Although our diffusion-based architecture involves a larger number of parameters, the overall computational cost, measured by per-frame latency and FLOPs, remains highly competitive. This demonstrates the effectiveness of our latent-space compression and the SDM with one-step sampling, which substantially reduce the computational overhead while preserving generative capability. This design not only lowers latency but also ensures stable temporal propagation.

6. Conclusion

SDMFusion that enables fast, high-fidelity infrared-visible video fusion. It leverages a one-step diffusion sampler for speed and a temporal adapter with memory for coherence. This approach effectively eliminates flickering artifacts while preserving both thermal and texture details. **Broader Impact.** This technology enhances perception in challenging conditions, benefiting day-night surveillance and autonomous navigation. We emphasize the need for ethical guidelines to mitigate potential privacy risks from enhanced surveillance capabilities, ensuring its responsible deployment for societal benefit.

Acknowledgments

This work was partially supported by the China Postdoctoral Science Foundation (2023M730741), the National Natural Science Foundation of China (No.62302078, No.62372080), and the Key Research and Development Program of Liaoning Province under Grant (No.2023JH26/10200014).

References

[1] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A Unified Solution to Video Fusion: From multi-frame learning to benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 3, 5, 6

[2] Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3250–3259, 2024. 1

[3] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24966–24976, 2024. 1

[4] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, 2022. 1

[5] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25912–25921, 2024. 1, 2

[6] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19248–19257, 2024. 1

[7] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma, and Boxin Shi. EventAid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6959–6973, 2025. 1

[8] Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. FreeFusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025. 1

[9] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8115–8124, 2023. 1

[10] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8082–8093, 2023. 1, 2, 3

[11] Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. RCVS: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*, 26:11031–11043, 2024. 2, 3, 6

[12] Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. TemCoCo: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14326–14335, 2025. 2, 3, 5, 6

[13] Hao Zhang, Lei Cao, and Jaiyi Ma. Text-DiFuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. In *Advances in Neural Information Processing Systems*, pages 39552–39572, 2024. 2, 3

[14] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27026–27035, 2024. 2

[15] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23174–23184, 2025. 2

[16] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and Visible Image Fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025. 2

[17] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023. 2

[18] Hui Li and Xiao-Jun Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 2

[19] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 976–976, 2021. 2

[20] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 2

[21] Ludan Sun, Kai Zhang, Feng Zhang, Wenbo Wan, and Jiande Sun. Deep rank-N decomposition network for image fusion. *IEEE Transactions on Multimedia*, 26:7335–7348, 2024. 2

[22] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916, 2023. 2, 5, 6

[23] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 2

[24] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 2

[25] Yuan Gao, Shiwei Ma, and Jingjing Liu. DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):549–561, 2023. 2

[26] Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-Fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32:5705–5720, 2023. 2

[27] Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-DiFuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 2, 5, 6

[28] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024. 2

[29] Zhengqiang Zhang, Ruihuang Li, Shi Guo, Yang Cao, and Lei Zhang. TMP: Temporal motion propagation for online video super-resolution. *IEEE Transactions on Image Processing*, 33:5014–5028, 2024. 2

[30] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-aware spatio-temporal sparse transformer for video deblurring. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3616–3626, 2024. 3

[31] Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution transformer with masked interintra-frame attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25399–25408, 2024. 3

[32] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6531–6539, 2024. 3

[33] Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. OmniFuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025. 3

[34] Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. LUT-Fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14559–14568, 2025. 3

[35] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023. 5, 6

[36] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024. 5, 6

[37] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. DCEvo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2226–2235, 2025. 5, 6

[38] Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One Model for ALL: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28102–28112, 2025. 5, 6

[39] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every SAM Drop Counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17882–17891, 2025. 5, 6

[40] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729, 2017. 5

[41] Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. VideoFusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025. 6

[42] Chen Sun, Xinyu Wang, Shenghua Fan, Xiaobing Dai, Yuting Wan, Xiao Jiang, Zengliang Zhu, and Yanfei Zhong. NOT-156: Night object tracking using low-light and thermal infrared: From multimodal common-aperture camera to benchmark datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. 6

[43] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. 2022. 8

[44] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 8

[45] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 8