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.

VideoFusion: A Spatio-Temporal Collaborative Network for Multi-modal Video Fusion

Linfeng Tang[1], Yeda Wang[1], Meiqi Gong[1], Zizhuo Li[1], Yuxin Deng[1]
Xunpeng Yi[1], Chunyu Li[1], Han Xu[2], Hao Zhang[1], Jiayi Ma[1]*
[1] Wuhan University, China [2]Southeast University, China
linfeng0419@gmail.com, jyma2010@gmail.com

Abstract

Compared to images, videos better reflect real-world acquisition and possess valuable temporal cues. However, existing multi-sensor fusion research predominantly integrates complementary context from multiple images rather than videos due to the scarcity of large-scale multi-sensor video datasets, limiting research in video fusion and the inherent difficulty of jointly modeling spatial and temporal dependencies in a unified framework. To this end, we construct ***M3SVD****, a benchmark dataset with* 220 *temporally synchronized and spatially registered infrared-visible videos comprising* 153,797 *frames, bridging the data gap. Secondly, we propose* ***VideoFusion****, a multi-modal video fusion model that exploits cross-modal complementarity and temporal dynamics to generate spatio-temporally coherent videos from multi-modal inputs. Specifically, 1) a differential reinforcement module is developed for cross-modal information interaction and enhancement, 2) a complete modality-guided fusion strategy is employed to adaptively integrate multi-modal features, and 3) a bi-temporal co-attention mechanism is devised to dynamically aggregate forward-backward temporal contexts to reinforce cross-frame feature representations. Experiments reveal that* ***VideoFusion*** *outperforms existing image-oriented fusion paradigms in sequences, effectively mitigating temporal inconsistency and interference. Project and M3SVD:* https://github.com/Linfeng-Tang/VideoFusion.

1. Introduction

Single-type sensors capture information from a limited perspective, making it challenging to comprehensively characterize imaging scenarios [51]. For instance, visible sensors rely on reflected light to capture detailed textures but are vulnerable to environmental factors. In contrast, infrared sensors leverage thermal radiation to highlight salient targets effectively, yet they lack fine-grained texture representation. Consequently, multi-sensor fusion, which integrates complementary information from diverse sensors to overcome the limitations of single-type sensors, has garnered considerable attention [9]. Among these, infrared and visible sensor fusion has emerged as a prominent research focus, demonstrating great potential in military detection [25], security surveillance [56], assisted driving [1].

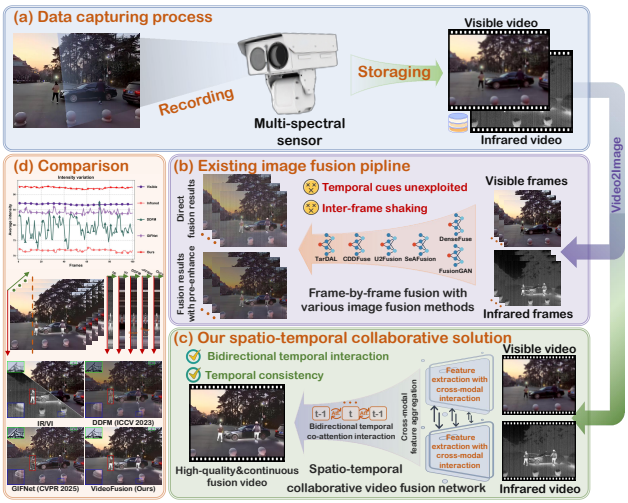


Figure 1. Image-oriented fusion vs. video fusion.

Earlier multi-sensor fusion studies primarily focused on image fusion, introducing various sophisticated networks, such as auto-encoders [11], convolutional neural networks [19], generative adversarial networks [22], Transformers [24], and diffusion models [60], in pursuit of superior visual perception. Moreover, some semantic-driven methods [33], considering semantic demands of downstream tasks, are proposed to reinforce the bond between image fusion and high-level vision tasks [17, 58]. Besides, some joint registration and fusion proposals [44], and

*Corresponding author.

degradation-robust solutions [35, 45] are also designed to cope with interference factors encountered in the imaging process. Particularly, degradation-robust methods leverage cross-modal complementary context to suppress degradation and directly generate high-quality fusion results [52], expanding the application scope of image fusion.

Although these methods achieve satisfactory performance, they neglect a crucial aspect that in practical applications, multi-modal sensors typically capture continuous video sequences rather than independent static frames, as illustrated in Fig. 1 (a). However, existing multi-sensor fusion methods are tailored for static images, primarily due to *the lack of large-scale multi-modal video datasets*. Consequently, they exploit cross-modal complementary features while overlooking inherent temporal dependencies in videos. It is worth noting that naively extending image fusion to frame-wise video fusion (Fig. 1 (b)) ignores inter-frame complementarity and introduces temporal incoherence, leading to inter-frame flickering artifacts, as shown in Fig. 1 (d). Furthermore, recent advances in video restoration have demonstrated that temporal cues play a crucial role in mitigating visual degradations such as deblurring [28], deraining [15], and low-light enhancement [21].

In this work, we first construct **M3SVD**, a multi-modal multi-scene video dataset with 220 temporally synchronized and spatially aligned infrared-visible videos totaling $153,797$ frames. Moreover, we propose a spatio-temporal collaborative video fusion network (**VideoFusion**), as presented in Fig. 1(c), to exploit temporal cues and cross-modal complementarity for robust representation. On the one hand, existing studies indicate that different modalities are typically complementary but also contain redundant information [47]. Thus, we design a cross-modal differential reinforcement module that exploits the complementary yet non-redundant information across modalities to achieve cross-modal information interaction and enhancement. Additionally, we develop a complete modality-guided fusion module to integrate enhanced features, which employs the sum of comprehensive infrared and visible features as a query to aggregate cross-modal complementarity. On the other hand, recognizing that cross-frame features offer dynamic cues to enrich scene representation, we devise a bidirectional temporal co-attention mechanism to integrate informative features from neighboring frames in both forward and backward directions into the current frame. By combining forward and backward attention, it effectively leverages temporal cues. To this end, the proposed method leverages complementary information across modalities and exploits temporal cues across frames to produce high-quality fused videos from degraded inputs, as illustrated in Fig. 1(d). In summary, our main contributions are as follows:

- A multi-modal multi-scene video dataset (M3SVD), comprising 220 temporally synchronized and spatially registered infrared-visible videos with $153,797$ frames, is constructed as the large-scale benchmark for video fusion, restoration, registration, and related areas.
- We propose a spatio-temporal collaborative network for infrared and visible video fusion, pioneering the joint modeling of cross-modal complementarity and across-frame temporal cues, advancing video fusion tasks.
- We devise a bi-temporal co-attention mechanism with a variational consistency loss to exploit forward and backward temporal cues, coupled with a differential reinforcement module to harness cross-modal complementary context for information restoration and integration.

2. Related Work

2.1. Image Fusion

At an early stage, image fusion methods primarily focus on aggregating cross-modal complementary information to enhance visual perception. Elaborate network architectures, including CNN [43, 62], AE [11, 12], GAN [16, 52], Transformer [24, 59], and diffusion models [36, 60], along with loss functions, such as intensity [23], gradient [50], SSIM [24] and perceptual [57] losses, are employed to preserve source-consistent meaningful information, effectively enhancing human visual perception. In particular, considering semantic requirements of downstream tasks, researchers proposed semantic-driven algorithms that leverage semantic segmentation [17, 33, 53] or object detection [16, 58] to enhance semantic preservation in fusion networks. Moreover, integrated registration and fusion frameworks [13, 32, 44] are introduced to simultaneously achieve alignment and fusion, mitigating parallax and distortion in real-world imaging and reducing artifacts in fusion results. Furthermore, degradation-robust paradigms [35, 37, 48, 54] are proposed to counteract degradation interference in the imaging process. However, these algorithms only leverage limited information available in static images to enrich scene representation, thereby failing to fully harness the potential of temporal cues inherent in video sequences.

2.2. Video Fusion

A few studies have recently focused on video fusion, such as RCVS [42], saPIDFuse [38], and DFTP [7]. Specifically, Xie *et al.* developed a unified registration and fusion framework for video streams [42]. However, these methods either adopt frame-by-frame strategies or rely on handcrafted features for temporal modeling, thus limiting the exploitation of temporal cues and harming temporal consistency. Our concurrent works, including TemCoCo [6] and UniVF [61], attempt to directly aggregate video sequences to alleviate inter-frame jitter. However, they depend on DCN or optical flow for inter-frame information compensation. DCN estimates frame-to-frame offsets in an unsupervised manner,

Table 1. Comparison of aligned multi-modal datasets.

Datasets	Temporal	Video nums	Image pairs	Resolution	Challenging scenarios			
					Low-light	Over-exp.	Disguise	Occlusion
RoadScene	No	–	221	768×576	✓	✓	✗	✗
MSRS	No	–	1,444	640×480	✓	✓	✗	✗
M^{3}FD	No	–	4,200	1024×768	✓	✓	✗	✓
FMB	No	–	1,500	800×600	✓	✓	✗	✓
LLVIP	No	–	15,488	**1280×1024**	✓	✓	✗	✗
TNO	Yes	3	114	768×576	✓	✗	✗	✓
INO	Yes	15	12,695	328×254	✗	✓	✗	✗
HDO	Yes	24	7,500	640×480	✓	✓	✗	✗
M3SVD	Yes	**220**	**153,797**	640×480	✓	✓	✓	✓

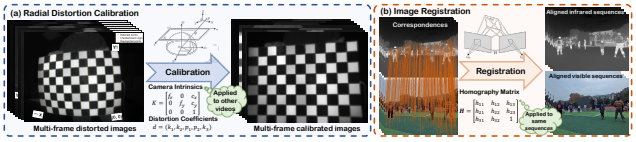


Figure 2. Schematic of image calibration and registration.

which may lead to instability, while optical flow networks, typically trained on visible images, struggle to generalize across multi-modal datasets. This work explores the potential of *attention mechanisms* to adaptively aggregate cross-modal and temporal complementary information, providing a more comprehensive and robust scene representation.

3. Multi-Modal Multi-Scene Video Dataset

As shown in Tab. 1, although multiple multi-modal datasets have been developed for image fusion tasks, most existing datasets (*e.g.*, RoadScene [43], MSRS [34], M^{3}FD [16], and FMB [18]) focus on static image pairs, lacking temporal coherence and limiting their applicability to dynamic scenarios. Only a small subset (*e.g.*, TNO [39], INO, and HDO [42]) incorporate temporal information, yet these video datasets remain scarce and critically underdeveloped. For instance, TNO has limited data, INO suffers from low resolution, and HDO exhibits poor imaging quality, with all covering only a narrow range of scenarios. Additional dataset comparisons are included in the **Suppl. Material**.

Therefore, we construct a multi-modal multi-scene video dataset (M3SVD) with a synchronized dual-spectral imaging system[1], to support standardized training and evaluation of multi-modal video fusion. The acquisition setup comprises: 1) an uncooled infrared sensor ($7.5 \sim 14\mu$m spectral response, 640×480 resolution, @30FPS), and 2) a visible CMOS sensor (1920×1080 resolution, @30FPS). As shown in Fig. 2, we first employ the *Matlab Camera Calibrator* with the checkerboard calibration plate to correct the radial distortion in both sensors. Second, due to the non-coincident optical centers and scale variations between two sensors, image registration is necessary. Given the 3.5cm inter-optical-axis distance between the sensors, the transformation between infrared and visible frames can be effectively approximated by a homography matrix $\mathbf{H}$ [14, 55]. To enhance the stability of registered videos, we estimate a single $\mathbf{H}$ for one video pair. We uniformly sample image pairs from multi-modal videos to establish precise correspondences using ReDFeat [5], XCP-Match [46], and MINIMA [30], followed by MAGSAC++ [2] to estimate $\mathbf{H}$. Finally, we register each infrared frame to the corresponding downsampled visible frame using $\mathbf{H}$, resulting in infrared and visible videos with a resolution of 640×480 at 30 FPS.

[1]https://www.magnity.com.cn/product/id/76.html

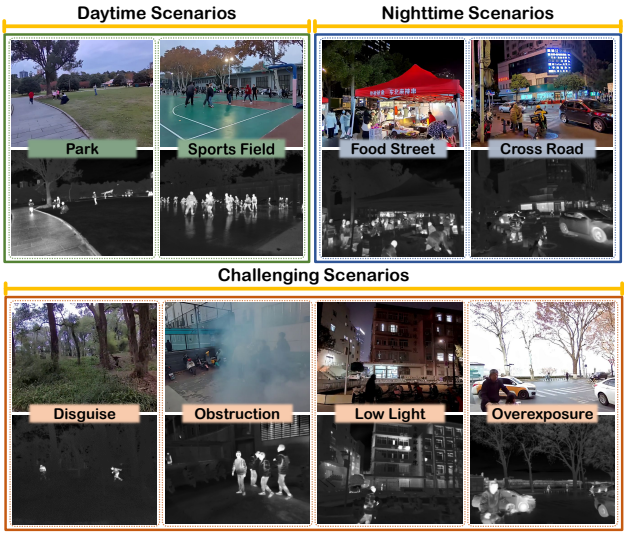


Figure 3. Visualization of various scenarios in M3SVD dataset.

As shown in Fig. 3, M3SVD encompasses daytime, nighttime, and challenging scenarios, covering diverse locations such as parks, lakes, sports fields, food streets, and crossroads. The challenging scenes involve typical multi-modal application surroundings such as disguise, occlusion, low light, and overexposure. In total, we collect 220 time-synchronized and spatially aligned multi-modal videos with $153,797$ frames, spanning 100 distinct scenes.

4. Methodology

4.1. Overview

Our workflow is illustrated in Fig. 4 (a). Given low-quality source videos $\{\tilde{V}_{ir}^i\}_{i=1}^T$ ($\tilde{V}_{ir}^i \in \mathbb{R}^{1\times H\times W}$) and $\{\tilde{V}_{vi}^i\}_{i=1}^T$ ($\tilde{V}_{vi}^i \in \mathbb{R}^{3\times H\times W}$), we propose a spatio-temporal collaborative network to generate a high-quality fused video $\{V_f^i\}_{i=1}^T$ ($V_f^i \in \mathbb{R}^{3\times H\times W}$), along with degradation-free infrared and visible videos $\{\hat{V}_{ir}^i\}_{i=1}^T$ and $\{\hat{V}_{vi}^i\}_{i=1}^T$. In the encoding stage, we first apply 3D convolution (Conv3D) to extract shallow temporal features $\mathcal{F}_x^1 \in \mathbb{R}^{T\times C_1\times H\times W}$, where $x \in \{ir, vi\}$. Then, a sequential architecture composed of downsampling layers, Conv3D, ResBlock, and a cross-modal differential reinforcement module (CmDRM) is utilized to extract multi-scale temporal features $\mathcal{F}_x^n \in \mathbb{R}^{T\times C_n\times H_n\times W_n}$

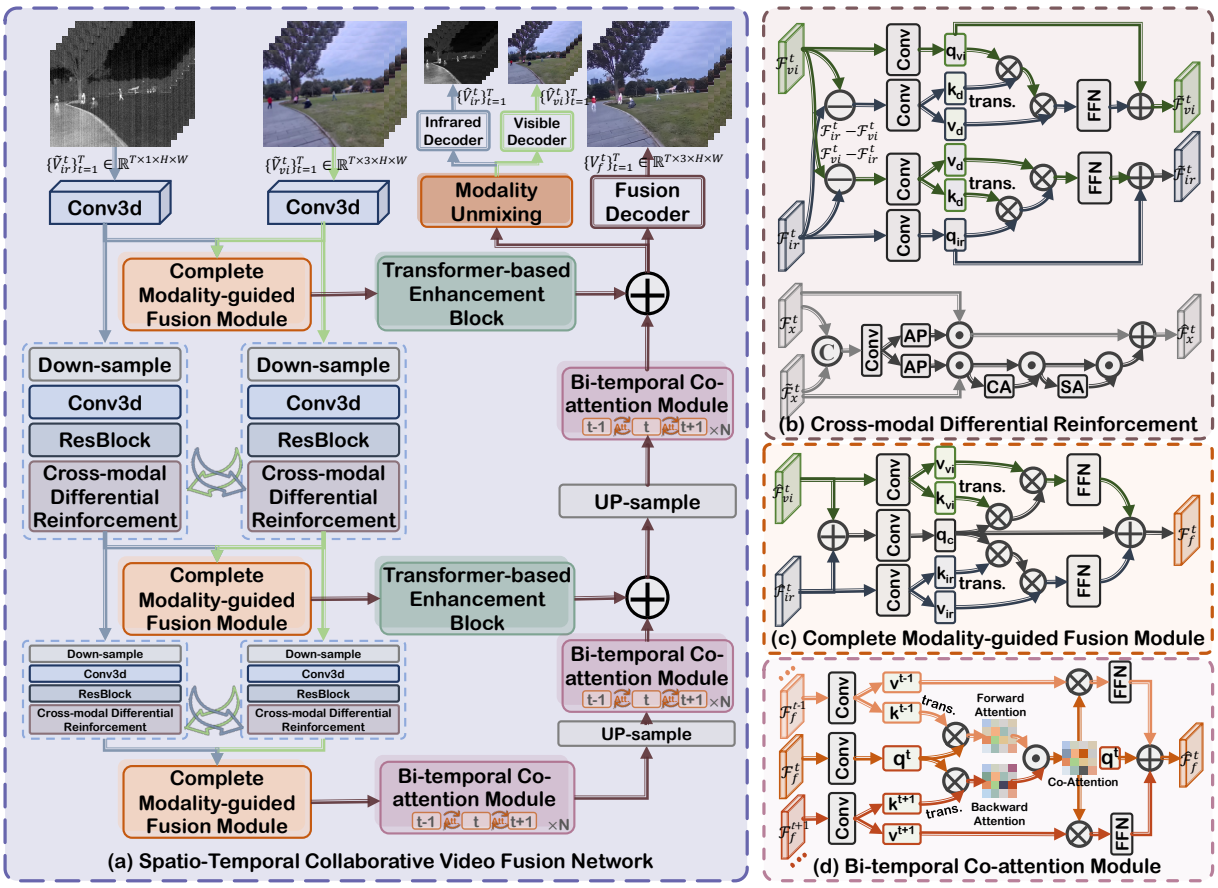


Figure 4. The overall framework of our spatio-temporal collaborative video fusion network.

for $n \in \{2, 3\}$. In particular, CmDRM enhances unimodal representations by integrating cross-modal differential information. Subsequently, complete modality-guided fusion modules are employed at multiple scales ($n \in \{1, 2, 3\}$) to aggregate cross-modal context and generate informative fusion features $\mathcal{F}_f^n$. To further enrich the feature representation, transformer-based enhancement blocks are introduced, followed by a bi-temporal co-attention module (BiCAM) in the decoding stage. BiCAM establishes dynamic temporal dependencies that go beyond Conv3D operations. Finally, a fusion decoder reconstructs the high-quality fusion video $\{V_f^i\}_{i=1}^T$, and a modality unmixing module with dedicated infrared and visible decoders yields $\{\hat{V}_{ir}^i\}_{i=1}^T$ and $\{\hat{V}_{vi}^i\}_{i=1}^T$. This design promotes degradation-invariant fusion while preserving cross-modal information.

4.2. Network Architectures

4.2.1. Cross-modal Differential Reinforcement Module

Given that features from various modalities introduce both complementary contexts and redundant information [34, 47], the cross-modal differential reinforcement module (CmDRM) first employs a cross-modal attention mechanism to aggregate differential information from complementary modalities, as illustrated in Fig. 4 (b). For conciseness, we omit scale indices (n), hence $\mathcal{F}_x^t \in \mathbb{R}^{C\times H\times W}(x \in \{ir, vi\})$ denote infrared or visible features of t-th frame. For instance, designating the visible modality as primary and the infrared as auxiliary, the complementary differential features for the primary modality are computed as $\mathcal{F}_d^t = \mathcal{F}_{ir}^t - \mathcal{F}_{vi}^t$, which captures unique information in the auxiliary modality. A 1×1 convolution is then applied to project $\mathcal{F}_d^t$ into keys (k_d) and values v_d, while mapping $\mathcal{F}_{vi}^t$ as queries (q_{vi}). The differentially reinforced features can be formulated as:

$$\tilde{\mathcal{F}}_{vi}^t = q_{vi} \oplus \text{FFN}\Big(softmax\Big(\frac{q_{vi}k_d^T}{\sqrt{d_k}}\Big)v_d\Big), \qquad (1)$$

where $\sqrt{d_k}$ is a scaling factor and FFN$(\cdot)$ denotes the feed-forward network. Furthermore, given the distinct contributions of primary features and differentially reinforced features to scene characterization, we devise a learnable contribution measurement to integrate both components while adaptively balancing their contributions. As shown in Fig. 4 (b), $\mathcal{F}_x^t$ and $\tilde{\mathcal{F}}_x^t$ are first concatenated along the chan-

nel dimension. The result is then processed through convolutional layers followed by average pooling to derive contribution metric scores $(w, \tilde{w})$. The weighted features are further refined through channel and spatial attention. The final cross-modal reinforced features are formulated as:

$$\begin{aligned} \tilde{\mathcal{F}}_{x_c}^t &= (\tilde{w} * \tilde{\mathcal{F}}_x^t) * \mathrm{CA}(\tilde{w} * \tilde{\mathcal{F}}_x^t), \\ \tilde{\mathcal{F}}_{x_s}^t &= \tilde{\mathcal{F}}_{x_c}^t * \mathrm{SA}(\tilde{\mathcal{F}}_{x_c}^t),\ \hat{\mathcal{F}}_x^t = w * \mathcal{F}_x^t \oplus \tilde{\mathcal{F}}_{x_s}^t, \end{aligned} \tag{2}$$

$\mathrm{CA}(\cdot)$ and $\mathrm{SA}(\cdot)$ here denote channel and spatial attention.

4.2.2. Complete Modality-guided Fusion Module

Although CmDRM enhances unimodal feature representations via cross-modal differential information, it remains insufficient for comprehensive scene characterization. To address this, we propose a complete modality-guided fusion (CMGF) module to aggregate complementary contexts, as shown in Fig. 4 (c). We hypothesize that simply summing features across modalities yields generic comprehensive features but lacks modality-specific expressiveness. Thus, we project comprehensive features $\mathcal{F}_c^t = \hat{\mathcal{F}}_{ir}^t + \hat{\mathcal{F}}_{vi}^t$ as public queries q_c to distill modality-specific information. The modality-specific features are mapped to corresponding keys (k_{ir}, k_{vi}) and values (v_{ir}, v_{vi}). The feature aggregation process can be formulated as:

$$\begin{aligned} \mathcal{F}_f^t = q_c \oplus \mathrm{FFN}\Big(softmax\Big(\frac{q_c k_{ir}^T}{\sqrt{d_k}}\Big) v_{ir}\Big) \\ \oplus \mathrm{FFN}\Big(softmax\Big(\frac{q_c k_{vi}^T}{\sqrt{d_k}}\Big) v_{vi}\Big). \end{aligned} \tag{3}$$

4.2.3. Bi-temporal Co-attention Module

Compared to frame-wise image fusion, video fusion better exploits inter-frame dependencies to suppress disturbances and maintain temporal coherence. However, solely relying on Conv3d inadequately unleashes the inherent potential of temporal cues. To this end, we propose a bi-temporal co-attention module (BiCAM) to establish dense cross-frame interactions via mutual attention, as shown in Fig. 4 (d). Given the current frame feature $\mathcal{F}_f^t$, we first establish interactions with neighboring frames through a standard multi-head cross-attention mechanism:

$$\mathcal{A}^{t-1} = softmax\Big(\frac{q^t k^{t-1T}}{\sqrt{d_k}}\Big), \mathcal{A}^{t+1} = softmax\Big(\frac{q^t k^{t+1T}}{\sqrt{d_k}}\Big), \tag{4}$$

where q^t is the shared query derived from the current frame feature $\mathcal{F}^t$. The forward key/value k^{t-1}/v^{t-1} and backward key/value k^{t+1}/v^{t+1} are projected from the previous frame feature $(\mathcal{F}_f^{t-1})$ and subsequent frame feature $(\mathcal{F}_f^{t+1})$. In particular, $q^t, k^{t-1}, k^{t+1}, v^{t-1}, v^{t+1} \in \mathbb{R}^{C_n \times H_n W_n}$. For boundary frames, we replicate the current frame as the neighbor: $\mathcal{F}_f^{t-1} = \mathcal{F}^t$ when $t = 1$ and $\mathcal{F}_f^{t+1} = \mathcal{F}^t$ when $t = T$ (thus $(k^0, v^0) = (k^1, v^1)$ or $(k^{T+1}, v^{T+1}) = (k^T, v^T)$). A co-attention mechanism is then introduced to enable bidirectional cross-temporal dynamic interactions:

$$\mathcal{A}_{co} = softmax(\mathcal{A}^{t-1} * \mathcal{A}^{t+1}). \tag{5}$$

Finally, we formulate cross-temporal aggregation as:

$$\hat{\mathcal{F}}_f^t = q^t \oplus \mathrm{FFN}(\mathcal{A}_{co} v^{t-1}) \oplus \mathrm{FFN}(\mathcal{A}_{co} v^{t+1}). \tag{6}$$

Note that we deploy N consecutive BiCAMs, *allowing each frame to assimilate complementary cues from adjacent frames and access long-range cross-temporal contexts using neighboring frames as mediators*. This is analogous to *the shifted window mechanism* in Swin Transformer [20].

4.2.4. Feature Enhancement and Modality Unmixing

The enhancement block employs the efficient transformer from Restormer [49] as its core operator. The fusion decoder and infrared/visible decoder share a homogeneous structure with this block to restore image representations. Moreover, modality unmixing is performed on fusion features via joint channel-spatial attention mechanisms [41].

4.3. Loss Functions

Adhering to the typical image fusion paradigm [24, 37, 48], our framework first constructs intensity loss $\mathcal{L}_{int}$, gradient loss $\mathcal{L}_{grad}$, and color loss $\mathcal{L}_{color}$ to effectively preserve discriminative information from multi-modal inputs. The definitions of $\mathcal{L}_{int}$, $\mathcal{L}_{grad}$, $\mathcal{L}_{color}$ are as follows:

$$\mathcal{L}_{int} = \tfrac{1}{HW} \textstyle\sum_{t=1}^{T} \|V_f^t - \max(V_{vi}^t, V_{ir}^t)\|_1, \tag{7}$$

$$\mathcal{L}_{grad} = \tfrac{1}{HW} \textstyle\sum_{t=1}^{T} \|\nabla V_f^t - \max(\nabla V_{vi}^t, \nabla V_{ir}^t)\|_1, \tag{8}$$

$$\mathcal{L}_{color} = \tfrac{1}{HW} \textstyle\sum_{t=1}^{T} \|\Phi_{\mathrm{CbCr}}(V_f^t) - \Phi_{\mathrm{CbCr}}(V_{vi}^t)\|_1, \tag{9}$$

where $\{V_{ir}^t, V_{vi}^t\}_{t=1}^T$ denotes high-quality infrared-visible videos, $\max(\cdot)$ selects salient targets and textures, $\|\cdot\|_1$ is the l_1-norm, ∇ denotes the Sobel operator, and $\Phi_{CbCr}(\cdot)$ converts RGB to CbCr. In particular, we use the Y channel of V_f^t and V_{vi}^t in $\mathcal{L}_{int}$ and $\mathcal{L}_{grad}$. Moreover, we introduce the scene fidelity loss $\mathcal{L}_{sf}$ to fully exploit the potential of modality unmixing and infrared/visible decoder, defined as:

$$\mathcal{L}_{sf} = \tfrac{1}{HW} \textstyle\sum_{x \in \{vi, ir\}} \sum_{t=1}^{T} (\|\hat{V}_x^t - V_x^t\|_1 + \|\nabla \hat{V}_x^t - \nabla V_x^t\|_1). \tag{10}$$

Besides, we assume that temporal variations of static backgrounds across consecutive frames are locally smooth, *i.e.*, $\mathbb{E}[\Delta V_{x_{\mathrm{bg}}}^t] \approx 0$, where $\Delta V_x^t = V_x^{t+1} - V_x^t$ and $x \in \{ir, vi\}$. Thus, enforcing $\|\Delta V_{f_{\mathrm{bg}}}^t - \Delta V_{x_{\mathrm{bg}}}^t\|_1 \to 0$ helps preserve temporal consistency in fused backgrounds. For dynamic objects, their motion-induced temporal variations should remain consistent with those in V_x. Therefore, we impose $\|\Delta V_{f_{\mathrm{ob}}}^t - \Delta V_{x_{\mathrm{ob}}}^t\|_1 \to 0$. Both can be unified as

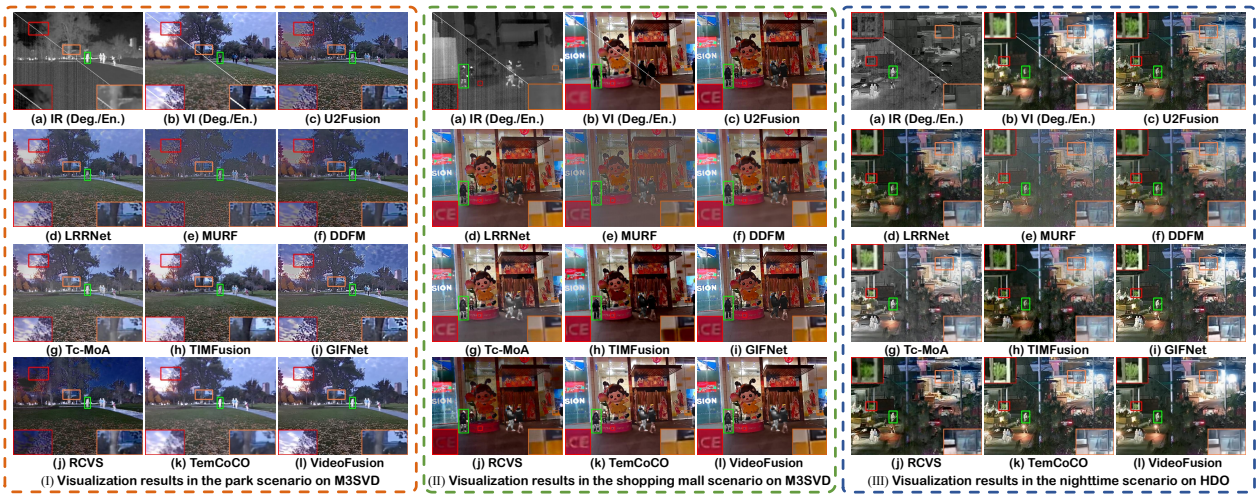


Figure 5. Qualitative comparison results on M3SVD and HDO datasets under degraded scenarios.

$\|\Delta V_f^t - \Delta V_x^t\|_1 \to 0$, and the same constraint is also applied to restored videos $\hat{V}_x$. Finally, we design a *variational consistency loss* $\mathcal{L}_{var}$ to prevent temporal flickering artifacts in fusion results and restored videos:

$$\mathcal{L}_{var} = \frac{1}{HW}\sum_x \sum_{t=1}^{T-1} (\underbrace{\|\Delta V_f^t - \Delta V_x^t\|_1}_{\mathcal{L}_{var}^{fu}} + \underbrace{\|\Delta \hat{V}_x^t - \Delta V_x^t\|_1}_{\mathcal{L}_{var}^{x}}). \quad (11)$$

Finally, the total loss of VideoFusion is formulated as:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_{int} \cdot \mathcal{L}_{int} + \lambda_{grad} \cdot \mathcal{L}_{grad} + \\ & \lambda_{color} \cdot \mathcal{L}_{color} + \lambda_{sf} \cdot \mathcal{L}_{sf} + \lambda_{var} \cdot \mathcal{L}_{var},\end{aligned} \quad (12)$$

where λ_{int}, λ_{grad}, λ_{color}, λ_{sf}, and λ_{var} are hyper-parameters for balancing various losses.

5. Experiments

5.1. Configurations and Implementation Details

We set $T = 7$ during training due to GPU memory limitations and $T = 25$ for testing. Other key parameters are set as $[N_1, N_2, N_3] = [2, 2, 4]$ and $[C_1, C_2, C_3] = [32, 64, 128]$. The hyper-parameters balancing various losses are empirically set as $\lambda_{grad} = 1$, $\lambda_{sf} = 10$, $\lambda_{int} = 15$, $\lambda_{color} = 100$, $\lambda_{var} = 100$. We train our VideoFusion for 20 epochs using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and an initial learning rate of 1×10^{-4} decayed to 1×10^{-5} via cosine annealing. Training is conducted on our M3SVD dataset with 200 videos, where visible videos are degraded by Gaussian blur (kernel size 15, standard deviation in $[0.9, 2.1]$) and infrared videos by stripe noise generated via a physics-inspired model [3]. Our method is implemented in PyTorch. All experiments are conducted on NVIDIA RTX 4090 GPUs and a 2.50 GHz Intel Xeon Platinum 8180 CPU.

We compare fusion performance with SOTA image fusion methods, including U2Fusion [43], LRRNet [12], MURF [44], DDFM [60], TC-MoA [62], TIMFusion [19], and GIFNet [4], as well as video fusion approaches, *i.e.* RCVS [42] and TemCoCo [6], under degraded and normal scenarios. We also employ SOTA video deblurring (*i.e.*, DSTNet [26]) and video denoising (*i.e.*, MDIVDNet [3]) algorithms to pre-enhance degraded videos for a fair comparison. Both MDIVDNet and DSTNet are retrained on our dataset with their default configurations. Six metrics are employed to comprehensively evaluate fusion methods from two perspectives, *i.e.,* fusion quality and temporal consistency. Particularly, EN [31], MI [27], SD [29], SSIM [40], and VIF [8] measure fusion quality, while flowD [6] assesses temporal coherence between frames.

5.2. Fusion Performance Comparison

Qualitative results under degraded scenarios on M3SVD and HDO datasets are presented in Fig. 5. As illustrated in Fig. 5 (I)-(II), U2Fusion, LRRNet, MURF, DDFM, TIMFusion, and GIFNet struggle to preserve prominent targets in the infrared videos, with MURF and TIMFusion introducing artifacts. Among video-based schemes, RCVS introduces more artifacts and reduces the overall brightness distribution of fusion results. TemCoCo, on the other hand, suffers from significant blur, although it attempts to leverage temporal cues to mitigate degradation. This suggests that its information aggregation and restoration modules still have inherent limitations. In contrast, our VideoFusion effectively integrates complementary and temporal information, producing clearer and more comprehensive scene representations. In nighttime scenes from the HDO dataset, as shown in Fig. 5 (III), video fusion methods including TemCoCo and VideoFusion not only enhance details but also highlight significant targets, such as pedestrians.

Quantitative evaluations on 20 videos each from M3SVD and HDO dataset are shown in Tab. 2. Video-

Table 2. Quantitative comparison on the M3SVD and HDO datasets under **degraded** scenarios. Each video in M3SVD and HDO contains 200 and 150 frames, respectively. The best and second-best results are highlighted in Red and Purple, respectively.

Methods	Multi-Modal Multi-Scene Video Dataset (M3SVD)						High-quality Dual-Optical Dataset (HDO)					
	EN ↑	MI ↑	SD ↑	SSIM ↑	VIF ↑	flowD ↓	EN ↑	MI ↑	SD ↑	SSIM ↑	VIF ↑	flowD ↓
U2Fusion [43]	6.904	2.490	35.731	0.600	0.439	6.547	7.015	1.993	48.473	0.560	0.385	8.723
LRRNet [12]	6.889	3.120	38.251	0.609	0.452	6.874	6.787	2.026	42.415	0.523	0.325	8.224
MURF [44]	6.185	1.825	22.731	0.543	0.271	6.840	6.599	1.718	32.640	0.532	0.324	9.410
DDFM [60]	6.750	2.656	32.000	0.609	0.448	6.778	7.019	2.017	42.614	0.566	0.381	7.546
TC-MOA [62]	7.095	2.800	42.412	0.593	0.516	5.102	**7.431**	2.089	52.766	0.543	0.460	8.846
TIMFusion [19]	7.063	3.015	50.824	0.580	0.409	5.890	6.639	1.998	53.146	0.481	0.339	7.221
GIFNet [4]	7.042	2.548	44.207	0.580	0.389	6.294	7.106	1.901	**59.287**	0.53	0.349	9.318
RCVS [42]	6.551	1.979	37.468	0.585	0.471	6.443	7.229	1.992	56.277	0.519	0.415	9.677
TemCoCo [6]	**7.174**	3.548	50.421	0.597	0.490	4.378	7.352	2.402	57.001	0.573	0.473	6.964
VideoFusion	7.167	**4.008**	**52.465**	**0.632**	**0.526**	**3.294**	7.288	**2.682**	57.024	**0.582**	**0.475**	**6.098**

Table 3. Quantitative comparison results on the M3SVD dataset under **normal** scenarios, where each video contains 200 frames.

Methods	EN ↑	MI ↑	SD ↑	SSIM ↑	VIF ↑	flowD ↓
U2Fusion [43]	6.935	2.659	36.316	0.614	0.505	6.615
LRRNet [12]	6.876	3.219	37.736	0.618	0.488	7.473
MURF [44]	6.209	1.915	22.844	0.546	0.310	7.373
DDFM [60]	6.777	2.642	31.960	0.593	0.462	7.064
TC-MOA [62]	7.109	2.894	42.711	0.602	0.577	5.305
TIMFusion [19]	7.085	3.083	51.637	0.590	0.433	6.295
GIFNet [4]	6.881	2.949	43.829	**0.646**	0.495	4.082
RCVS [42]	6.592	2.123	37.638	0.606	0.595	5.547
TemCoCo [6]	7.174	3.548	50.423	0.597	0.490	4.379
VideoFusion	**7.199**	**4.191**	**53.375**	**0.646**	**0.605**	**3.494**

Fusion achieves the highest MI and SSIM, indicating that our method effectively transfers complementary information from source videos to fusion results. Meanwhile, the superior VIF demonstrates that VideoFusion better aligns with human visual perception. Comparable EN and SD suggest that our results contain abundant information and high contrast. Among video-based alternatives, TemCoCo achieves comparable performance, whereas the frame-by-frame approach RCVS exhibits limited performance on these metrics, further underscoring the necessity of leveraging temporal sequence information. Additional results on normal M3SVD scenarios are reported in Tab. 3, where VideoFusion demonstrates superior performance across all metrics. This further validates its ability to fully exploit cross-modal complementary cues and temporal context to counteract interference and enhance scene characterization.

5.3. Temporal Consistency Comparison

As shown in Tabs. 2 and 3, VideoFusion achieves the lowest flowD on both M3SVD and HDO datasets, indicating its superior temporal stability and consistency. TemCoCo also attains relatively low flowD, further highlighting the significance of temporal modeling. In contrast, frame-by-frame fusion schemes exhibit higher flowD, implying noticeable temporal jitter in their fused videos. Furthermore, following [26], we visualize the temporal variations of different videos in Fig. 6, where frame-by-frame fusion methods exhibit evident flickering, especially in DDFM and LRRNet. In contrast, our method achieves temporally stable fusion, benefiting from BiCAM and variational consistency loss. The stable mean intensity variation further confirms the superior temporal consistency of our method. Additionally, Fig. 7 illustrates the temporal trajectories of various metrics across consecutive sequences, reinforcing this advantage.

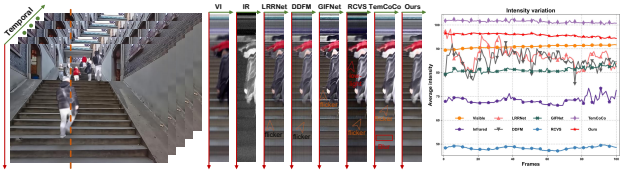


Figure 6. Temporal consistency comparison of source and fused videos. Following [26], column pixels (dotted line) are visualized and their average brightness variation is computed.

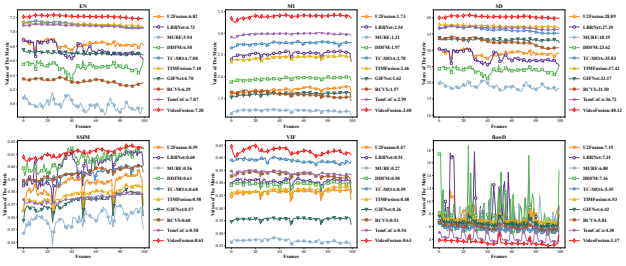


Figure 7. Temporal variation of metrics on sequences.

5.4. Extension Experiments

Object Tracking. Effective information enhancement and aggregation not only enhance visual perception, but also improve machine vision. Fig. 8 provides a qualitative as-

Table 4. Comparison of computational efficiency. * indicates methods incurring extra cost from preprocessing algorithms (*e.g.*, MDIVDNet and DSTNet) under degraded scenarios.

Methods	MDIVDNet	U2Fusion*	LRRNet*	MURF*	DDFM*	TC-MoA*
Parm. (M)	10.047	0.659	0.049	0.116	552.660	340.354
Flops (G)	240.15	405.2	14.17	31.49	5220.5	3932.09
Time (s)	0.033	0.077	0.011	0.195	34.502	0.183
Methods	**DSTNet**	**TIMFusion***	**GIFNet***	**RCVS***	**TemCoCo**	**VideoFusion**
Parm. (M)	7.448	1.232	3.291	0.670	19.183	6.743
Flops (G)	224.38	215.9	186.59	27.88	491.14	267.78
Time (s)	0.021	0.029	0.043	0.012	0.052	0.067



Figure 8. Visual comparison of object track on M3SVD dataset.

sessment of various fusion schemes for object tracking on M3SVD, where a pre-trained YOLO v11 [10] serves as the baseline tracker. On the one hand, the tracker detects more objects in our results, as our method effectively integrates cross-modal complementary information and temporal context, providing a more comprehensive scene representation. On the other hand, it predicts smoother trajectories on our results, benefiting from the improved temporal consistency.

Computational Efficiency. As shown in Tab. 4, VideoFusion achieves computational efficiency comparable to image-level fusion methods. Notably, in degraded scenarios, most algorithms incur additional computational overhead due to preprocessing steps, further demonstrating the efficiency advantage of our VideoFusion.

Ablation Studies. As shown in Fig. 9, we conduct a series of ablation studies to validate our key designs. Removing BiCAM or $\mathcal{L}_{int}$ weakens salient targets, while omitting CmDRM reduces information recovery. Besides, replacing CMGF with a simple summation causes severe distortion. Similarly, w/o $\mathcal{L}_{color}$ induces noticeable color distortion and artifacts, and w/o $\mathcal{L}_{grad}$ results in detailed texture loss. Moreover, as shown in Fig. 9 (j), removing BiCAM or $\mathcal{L}_{var}$

Table 5. Quantitative comparison of ablation studies.

Configs	EN ↑	MI ↑	SD ↑	SSIM ↑	VIF ↑	flowD ↓
w/o BiCAM	**7.250**	3.439	52.194	0.601	0.472	4.747
w/o CmDRM	6.953	3.557	48.439	0.612	0.510	3.728
w/o CMGF	7.046	2.099	**61.403**	0.366	0.233	7.029
w/o $\mathcal{L}_{grad}$	7.208	3.702	52.239	0.615	0.5	3.669
w/o $\mathcal{L}_{int}$	7.106	2.985	47.630	0.630	0.468	3.684
w/o $\mathcal{L}_{var}$	7.211	3.432	52.245	0.599	0.480	6.056
w/o $\mathcal{L}_{color}$	6.839	2.102	39.854	0.457	0.240	6.031
VideoFusion	7.167	**4.008**	52.465	**0.632**	**0.526**	**3.294**

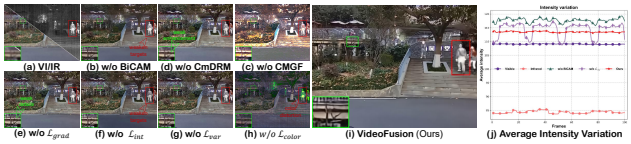


Figure 9. Visual comparison of ablation studies.

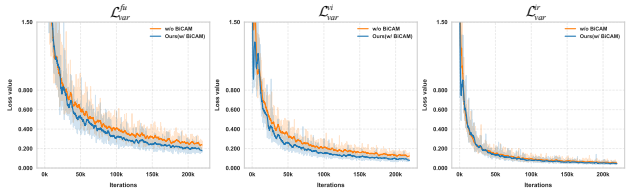


Figure 10. Training processes for BiCAM.

disrupts temporal consistency. Particularly, as presented in Fig. 10, w/o BiCAM affects the convergence of $\mathcal{L}_{var}$. Although Tab. 5 shows that the model without CMGF achieves optimal SD, this comes at the cost of introducing significant artifacts and distortions. In contrast, VideoFusion with these key components achieves more balanced overall performance while preserving temporal coherence.

6. Conclusion

This work has presented **M3SVD**, a multi-modal multi-scene video dataset, and **VideoFusion**, a spatio-temporal collaborative framework for multi-modal video fusion. On the one hand, a cross-modal differential reinforcement module has been devised for cross-modal information interaction and enhancement, while a complete modality-guided fusion strategy has been employed to integrate multi-modal features. On the other hand, a bi-temporal co-attention mechanism with a variational consistency loss has been designed to dynamically aggregate forward-backward temporal contexts, reinforcing cross-frame feature representations. Extensive experiments have demonstrated the superiority of VideoFusion over conventional image-based fusion paradigms in sequential scenarios, particularly in alleviating temporal inconsistency and interference.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (625B2135, U23B2050, 62506268, 62502085, 62276192).

References

[1] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023. 1

[2] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1304–1312, 2020. 3

[3] Lijing Cai, Xiangyu Dong, Kailai Zhou, and Xun Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024. 6

[4] 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 Conference on Computer Vision and Pattern Recognition*, pages 28102–28112, 2025. 6, 7

[5] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2022. 3

[6] 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 International Conference on Computer Vision*, pages 14326–14335, 2025. 2, 6, 7

[7] Xiaoming Guo, Fengbao Yang, and Linna Ji. Dftp: A mimic fusion method driven by difference feature time-phase perception for infrared and visible video. *IEEE Transactions on Instrumentation and Measurement*, 73:1–20, 2024. 2

[8] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. 6

[9] Shahid Karim, Geng Tong, Jinyang Li, Akeel Qadir, Umar Farooq, and Yiting Yu. Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90:185–217, 2023. 1

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

[11] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 1, 2

[12] 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. 2, 6, 7

[13] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[14] Zizhuo Li, Yifan Lu, Linfeng Tang, Shihua Zhang, and Jiayi Ma. Comatch: Dynamic covisibility-aware transformer for bilateral subpixel-level semi-dense image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 18521–18530, 2025. 3

[15] Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, Shunli Zhang, and Robby T Tan. Nightrain: Nighttime video deraining via adaptive-rain-removal and adaptive-correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3378–3385, 2024. 2

[16] 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 Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 2, 3

[17] 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 International Conference on Computer Vision*, pages 8115–8124, 2023. 1, 2

[18] 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 International Conference on Computer Vision*, pages 8115–8124, 2023. 3

[19] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024. 1, 6, 7

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[21] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Weigang Zhang, Hongxun Yao, and Qingming Huang. Unsupervised low-light video enhancement with spatial-temporal co-attention transformer. *IEEE Transactions on Image Processing*, 32:4701–4715, 2023. 2

[22] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. 1

[23] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70: 5009513, 2021. 2

[24] 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. 1, 2, 5

[25] Amanda C Muller and Sundaram Narayanan. Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, 10(2):137–149, 2009. 1

[26] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22191–22200, 2023. 6, 7

[27] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002. 6

[28] Chen Rao, Guangyuan Li, Zehua Lan, Jiakai Sun, Junsheng Luan, Wei Xing, Lei Zhao, Huaizhong Lin, Jianfeng Dong, and Dalong Zhang. Rethinking video deblurring with wavelet-aware dynamic transformer and diffusion model. In *Proceedings of the European Conference on Computer Vision*, pages 421–437. Springer, 2024. 2

[29] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement Science and Technology*, 8(4):355–375, 1997. 6

[30] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 23059–23068, 2025. 3

[31] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 6

[32] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022. 2

[33] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 1, 2

[34] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 3, 4

[35] Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the ACM International Conference on Multimedia*, pages 8546–8555, 2024. 2

[36] 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

[37] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion framework with language-vision degradation prompts. In *Advances in Neural Information Processing Systems*, 2025. 2, 5

[38] Xiaolin Tang, Jun Wang, and Linlu Dong. An infrared and visible light video fusion method based on chaos theory and pid control. *Infrared Physics & Technology*, 140:105422, 2024. 2

[39] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017. 3

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[41] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 5

[42] 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, 7

[43] 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, 3, 6, 7

[44] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 1, 2, 6, 7

[45] Han Xu, Xunpeng Yi, Chen Lu, Guangcan Liu, and Jiayi Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 34:5803–5818, 2025.

[46] Meng Yang, Fan Fan, Jun Huang, Yong Ma, Xiaoguang Mei, Zhanchuan Cai, and Jiayi Ma. Multimodal image matching based on cross-modality completion pre-training. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2206–2214, 2025. 3

[47] 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. 2, 4

[48] 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 Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024. 2, 5

[49] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 5

[50] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12797–12804, 2020. 2

[51] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[52] Hao Zhang, Linfeng Tang, Xinyu Xiang, Xuhui Zuo, and Jiayi Ma. Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26487–26496, 2024. 2

[53] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26974–26983, 2024. 2

[54] 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. 2

[55] Kaining Zhang and Jiayi Ma. Sparse-to-dense multimodal image registration via multi-task learning. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[56] Yue Zhang, Bin Song, Xiaojiang Du, and Mohsen Guizani. Vehicle tracking using surveillance with multimodal data fusion. *IEEE Transactions on Intelligent Transportation Systems*, 19(7):2353–2361, 2018. 1

[57] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118, 2020. 2

[58] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023. 1, 2

[59] 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 Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 2

[60] 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 International Conference on Computer Vision*, pages 8082–8093, 2023. 1, 2, 6, 7

[61] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. In *Advances in Neural Information Processing Systems*, 2025. 2

[62] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024. 2, 6, 7