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

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 [2]Southeast University
{linfeng0419,licy0089,zhpersonalbox,jyma2010}@gmail.com
{wangyeda,meiqigong,zizhuo_li,dyx_acuo,yixunpeng}@whu.edu.cn
xu_han@seu.edu.cn

Abstract

*Compared to images, videos better align with real-world acquisition scenarios and possess valuable temporal cues. However, existing multi-sensor fusion research predominantly integrates complementary context from multiple images rather than videos. This primarily stems from two factors: 1) the scarcity of large-scale multi-sensor video datasets, limiting research in video fusion, and 2) the inherent difficulty of jointly modeling spatial and temporal dependencies in a unified framework. This paper proactively compensates for the dilemmas. First, we construct **M3SVD**, a benchmark dataset with* 220 *temporally synchronized and spatially registered infrared-visible video pairs comprising* 153,797 *frames, filling the data gap for the video fusion community. Secondly, we propose **VideoFusion**, a multi-modal video fusion model that fully exploits cross-modal complementarity and temporal dynamics to generate spatio-temporally coherent videos from (potentially degraded) 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. Extensive experiments reveal that **VideoFusion** outperforms existing image-oriented fusion paradigms in sequential scenarios, effectively mitigating temporal inconsistency and interference.*

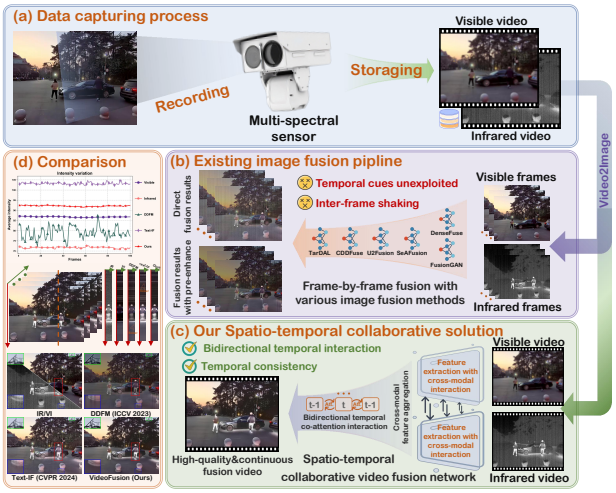

Figure 1. Image-oriented fusion vs. video fusion.

1. Introduction

Single-type sensors can only capture information from a specific perspective, making it difficult to comprehensively characterize the imaging scenarios [48]. For instance, visible sensors rely on reflected light for imaging and can capture texture details of objects, but they are vulnerable to environmental factors. On the other hand, infrared sensors leverage thermal radiation for imaging, effectively highlighting salient targets, but they lack the ability to represent fine-grained textures. Thus, multi-sensor fusion, aiming to integrate complementary information from various sensors to overcome the limitations of single-type sensors, has attracted significant attention [11, 52]. Among these, infrared and visible image fusion has emerged as a prominent research focus, demonstrating great potential in military detection [28], security surveillance [53], assisted driving [1], and scene understanding [8, 51].

Earlier image fusion methods introduced various sophis-

*Corresponding author.

ticated network architectures, such as auto-encoders [15], convolutional neural networks [23], generative adversarial networks [25], Transformers [27], and diffusion models [56], in pursuit of superior visual perception. Moreover, some semantic-driven approaches [33], considering semantic demands of downstream tasks, are proposed to reinforce the bond between image fusion and high-level vision tasks, *e.g.,* semantic segmentation [21] and object detection [20, 55]. Besides, some joint registration and fusion proposals [32, 42], as well as degradation-robust solutions [36, 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 interference and directly generate high-quality fusion results [49], significantly expanding the application scope of image fusion.

Although these methods achieve satisfactory fusion performance, they neglect a crucial aspect that in practical applications, multi-modal sensors typically capture video sequences instead of static frames, as illustrated in Fig. 1(a). However, existing multi-sensor fusion frameworks are tailored for static images, primarily due to the lack of large-scale multi-modal video dataset. Specifically, they exploit cross-modal complementary features while overlooking inherent temporal dependencies in video sequences. On the one hand, naively extending image fusion methods to frame-wise video fusion (Fig. 1(b)) not only ignores inter-frame contextual complementarity but also introduces temporal incoherence, leading to inter-frame flickering artifacts, as shown in Fig. 1(d). This undermines both perceptual quality and performance in downstream tasks. On the other hand, numerous video restoration algorithms demonstrate that temporal cues within videos often play an active role in combating degradations such as deblurring [29, 30], deraining [4, 19], and low-light enhancement [6, 24].

In this work, we first construct the multi-modal multi-scene video dataset (**M3SVD**), a comprehensive benchmark comprising 220 temporally synchronized and spatially aligned infrared-visible video pairs with $153,797$ frames. Furthermore, we propose a spatio-temporal collaborative video fusion network (**VideoFusion**), as presented in Fig. 1(c), to fully exploit the temporal cues in video sequences and cross-modal complementarity for a comprehensive description of imaging scenarios. On the one hand, existing studies indicate that different modalities are typically complementary but also contain redundant information [34, 43]. Thus, we design a cross-modal differential reinforcement module that exploits the complementary yet non-redundant differential information across modalities to achieve cross-modal information interaction and restoration. 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, considering that cross-frame features can provide dynamic cues to enhance scene descriptions, we devise a bidirectional temporal co-attention mechanism to integrate informative features from neighboring temporal frames in both forward and backward directions into the current frame. Specifically, it combines forward attention (between the current and previous frames) with backward attention (between the current and next frames), thereby better leveraging temporal cues for comprehensive cross-temporal perception and maintaining temporal consistency. To this end, the proposed method not only leverages complementary information across modalities but also utilizes temporal cues across frames to generate high-quality fused videos from degraded inputs, as shown in Fig. 1 (d). To sum up, 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, and related areas.
- We propose a groundbreaking spatio-temporal collaborative network for infrared and visible video fusion and restoration, 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 robust information restoration and integration.

2. Related Work

Image Fusion. At an early stage, visual-oriented image fusion methods primarily focus on aggregating cross-modal complementary information to enhance visual perception. Elaborate network architectures, including CNN [41, 58], AE [15, 16], GAN [20, 49], Transformer [27, 38], and diffusion models [7, 56], along with loss functions, such as intensity [26], gradient [47], SSIM [27] and perceptual [54] 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 [21, 33, 50] or object detection [20, 55] to enhance semantic preservation in fusion networks. Additionally, integrated registration and fusion frameworks [17, 32, 42] are introduced to simultaneously address image alignment and fusion, mitigating parallax and distortion in real-world imaging and reducing artifacts in fusion results. Furthermore, degradation-robust paradigms are proposed, such as DIVFusion [35], DDBF [49], DRMF [36], and Text-IF [45], to counteract

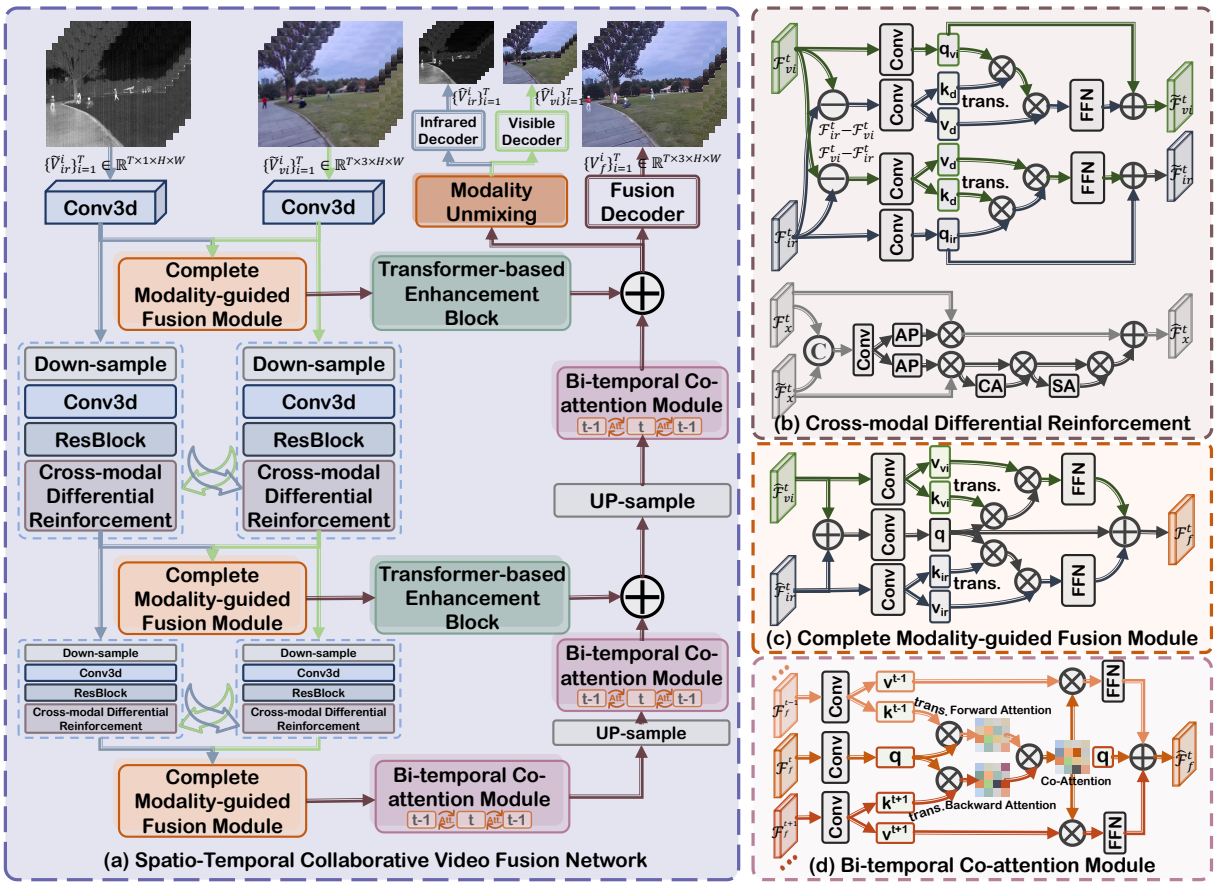

Figure 2. The overall framework of our spatio-temporal collaborative video fusion network.

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. Notably, although Xie *et al.* developed a unified registration and fusion framework for video streams [40], they still adopted a frame-by-frame approach to aggregate complementary information.

Video Restoration. The potential of temporal contexts has been successfully exploited in various restoration tasks, such as deblurring [29, 30], deraining [4, 19], denoising [3, 31], super-resolution [44, 57], and low-light enhancement [6, 24]. In particular, Pan *et al.* proposed a deep discriminative spatial and temporal network to facilitate the spatial and temporal feature exploration for better video deblurring [29]. Cai *et al.* devised a multi-domain infrared video denoising network that integrates temporal, spatial, and frequency-domain features to restore high-fidelity infrared videos [3]. Additionally, researchers also explored the potential of cross-modal information (*e.g.*, event [13, 14] and depth [18] data) in video restoration tasks, achieving impressive results. In this work, we attempt to combine cross-modal complementary information with temporal context to provide a more comprehensive and robust scene representation.

3. Methodology

3.1. Overview

Our workflow is illustrated in Fig. 2 (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 U-shaped spatio-temporal collaborative fusion network to generate the high-quality fusion video $\{V_f^i\}_{i=1}^T$ ($V_f^i \in \mathbb{R}^{3\times H\times W}$) and degradation-free infrared/visible videos $\{\hat{V}_{ir}^i\}_{i=1}^T$ and $\{\hat{V}_{vi}^i\}_{i=1}^T$. In the encoding stage, we first utilize the 3D convolutional module (Conv3D) to extract shallow temporal features $\mathcal{F}_x^n \in \mathbb{R}^{T\times C_n\times H\times W}$ where $x \in \{ir, vi\}$ and $n = 1$. Furthermore, a sequential architecture comprising downsampling layer, Conv3D, residual block (ResBlock), and cross-modal differential reinforcement module (CmDRM) is deployed to extract multi-scale temporal features $\mathcal{F}_x^n \in \mathbb{R}^{T\times C_n\times H_n\times W_n}$, where $x \in \{ir, vi\}$ and $n \in \{2, 3\}$. Particularly, CmDRM enhances unimodal representations by adaptively integrating complementary cross-modal dif-

ferential information during feature extraction. After that, complete modality-guided fusion modules are employed to aggregate cross-modal complementary context and generate informative fusion features $\mathcal{F}_f^n \in \mathbb{R}^{T\times C_n\times H_n\times W_n}$, where $n \in \{1,2,3\}$, at various scales. To strengthen feature representation, we introduce transformer-based enhancement blocks followed by a bi-temporal co-attention module (BiCAM) in the decoding phase, which establishes dynamic temporal dependencies beyond Conv3D-based interactions. Finally, a fusion decoder reconstructs high-quality video $\{V_f^i\}_{i=1}^T$, while a modality unmixing module with dedicated infrared/visible decoders yields restoration outputs ($\{\hat{V}_{ir}^i\}_{i=1}^T$ and $\{\hat{V}_{vi}^i\}_{i=1}^T$. This design enforces fused features to maintain degradation-invariant properties while preserving comprehensive multi-modal contexts.

3.2. Network Architectures

As mentioned above, we devise critical modules to harness cross-modal complementarity and dynamic temporal cues for comprehensive and interference-robust scene representation. This section outlines their architecture.

3.2.1. Cross-modal Differential Reinforcement Module

Given that features of complementary modalities introduce both complementary contexts and redundant information [34, 43], the cross-modal differential reinforcement module (CmDRM) first employs the cross-modal attention mechanism to aggregate differential information from complementary modalities, as illustrated in Fig. 2(b). For conciseness, we do not differentiate 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 means unique information in the auxiliary modality. Then, we apply a simple 1×1 convolution to project $\mathcal{F}_d^t$ into keys (k_d) and values v_d, while mapping F_{vi} as queries (q_{vi}). The differentially reinforced features can be formulated as:

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

where $\sqrt{d_k}$ is a scaling factor and FFN$(\cdot)$ is the feed-forward network. Furthermore, given the distinct contributions of primary features and differentially reinforced features to scene characterization, we elaborate a learnable contribution metric mechanism to integrate both components while adaptively balancing their contributions. As shown in Fig. 2(b), $\mathcal{F}_x^t$ and $\tilde{\mathcal{F}}_x^t$ are first concatenated along the channel dimension. The result is then processed through convolutional layers followed by average pooling to derive contribution metric scores ($w, \tilde{w}$). The weighted differentially reinforced features are further adjusted through channel and spatial attention mechanisms. The final cross-modal reinforced features are formulated as:

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

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

3.2.2. Complete Modality-guided Fusion Module

Although CmDRM enhances unimodal feature representations by leveraging cross-modal differential information, it fails to achieve comprehensive scenario characterization. To address this limitation, we devise a complete modality-guided fusion (CMGF) module to aggregate complementary contexts, as presented in Fig. 2(c). Specifically, we hypothesize that simply summing features across various modalities yields generic comprehensive features but lacks modality-specific expressiveness. Therefore, we project the comprehensive features $\mathcal{F}_c^t = \hat{\mathcal{F}}_{ir}^t + \hat{\mathcal{F}}_{vi}^t$ as public queries q_c to further distill modality-specific information from both infrared and visible modalities. Consequently, the infrared and visible features are mapped into their corresponding keys (k_{ir}, k_{vi}) and values (v_{ir}, v_{vi}). Thus, the feature aggregation process can be formulated as:

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

3.2.3. Bi-temporal Co-attention Module

Compared to frame-wise image fusion, video fusion can effectively exploit 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 through mutual attention mechanisms, as shown in Fig. 2(d). Given the current frame feature $\mathcal{F}_f^t$, we first establish the interaction of neighboring frames with the mutual attention mechanism:

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

where q^t is the shared query derived from the current frame feature $\mathcal{F}^t$. The forward key k^{t-1} and backward key k^{t+1} are projected from the previous frame feature ($\mathcal{F}_f^{t-1}$) and subsequent frame feature($\mathcal{F}_f^{t+1}$), respectively. We further introduce a co-attention mechanism to enable bidirectional cross-temporal dynamic interactions:

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

Finally, we formulate cross-temporal aggregation as:

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

Table 1. Comparison of different aligned multi-modal datasets.

Datasets	Temporal	Video nums	Image pairs	Resolution	Challenging scenarios			
					Low-light	Over-exp.	Disguise	Occlusion
RoadScene [41]	No	–	221	768×576	✓	✓	✗	✗
MSRS [34]	No	–	1,444	640×480	✓	✓	✗	✗
M^{3}FD [20]	No	–	4,200	1024×768	✓	✓	✗	✓
FMB [22]	No	–	1,500	800×600	✓	✓	✗	✓
LLVIP [9]	No	–	15,488	**1280×720**	✓	✓	✗	✗
TNO [37]	Yes	3	114	768×576	✓	✗	✗	✓
INO	Yes	15	12,695	328×254	✗	✓	✗	✗
HDO [40]	Yes	24	7,500	640×480	✓	✓	✗	✗
M3SVD	Yes	**220**	**153,797**	640×480	✓	✓	✓	✓

Note that we deploy N consecutive BiCAMs that enable each frame to both assimilate complementary information from adjacent frames and receive long-range cross-temporal contexts using neighboring frames as mediators.

3.2.4. Feature Enhancement and Modality Unmixing

The transformer-based enhancement block employs the efficient transformer of Restormer [46] as its core operator. The fusion decoder and infrared/visible decoder adopt a structurally homogeneous architecture with this block to restore image representations. Moreover, we implement modality unmixing from fusion features via joint channel-spatial attention mechanisms [39].

3.3. Loss Functions

Adhering to the typical image fusion paradigm [27, 50], our framework 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:

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

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

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

where $\{V_{ir}^t, V_{vi}^t\}_{t=1}^T$ denotes the high-quality infrared-visible video pair, $\max(\cdot)$ is the maximum selection for preserving salient targets and textures, $\|\cdot\|_1$ means the l_1-norm, ∇ is 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 $\mathcal{L}_{grad}$. Additionally, we introduce the scene fidelity loss $\mathcal{L}_{sf}$ to fully exploit the potential of modality unmixing and infrared/visible decoder, defined as:

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

Furthermore, we hypothesize that the temporal variation of static backgrounds in consecutive videos exhibits local smoothness, while the variation of dynamic objects in fusion (or restored source) videos should align with that in high-quality source videos. Formally:

$$\begin{cases} \mathbb{E}\left[\Delta V_{x_{\text{bg}}}^t\right] \to 0, & \text{(static backgrounds)} \\ \|\Delta V_{f_{\text{ob}}}^t - \Delta V_{x_{\text{ob}}}^t\|_1 \to 0, & \text{(dynamic objects)} \end{cases} \tag{11}$$

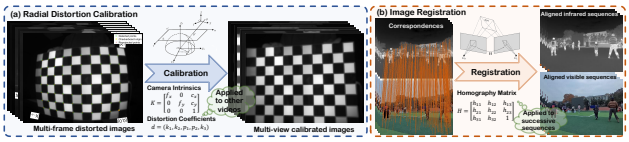

Figure 3. Schematic of image calibration and registration.

where $\Delta V_x^t = V_x^{t+1} - V_x^t$ and $x \in \{ir, vi\}$. Thus, we design a variational consistency loss $\mathcal{L}_{var}$ to prevent temporal flickering artifacts in fusion results and restored videos:

$$\mathcal{L}_{var} = \frac{1}{HW}\sum_{x}\sum_{t=1}^{T-1}(\underbrace{\|\Delta V_f^t - \Delta V_x^t\|_1}_{\mathcal{L}_{var}^{fu}} + \underbrace{\|\Delta\hat{V}_x^t - \Delta V_x^t\|_1}_{\mathcal{L}_{var}^{x}}). \tag{12}$$

Finally, the total loss of our VideoFusion is formulated as the weighted sum of aforementioned losses:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_{int}\cdot\mathcal{L}_{int} + \lambda_{grad}\cdot\mathcal{L}_{grad} + \\ & \lambda_{color}\cdot\mathcal{L}_{color} + \lambda_{sf}\cdot\mathcal{L}_{sf} + \lambda_{var}\cdot\mathcal{L}_{var},\end{aligned} \tag{13}$$

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

4. 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 [41], MSRS [34], M^{3}FD [20], and FMB [22]) focus on static image pairs, lacking temporal coherence and limiting their applicability to dynamic scenarios. Only a small subset (*e.g.*, TNO [37], INO[1], and HDO [40]) 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.

Therefore, we construct a multi-modal multi-scene video dataset (M3SVD) with a synchronized dual-spectral imaging system[2], 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 illustrated in Fig. 3, 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 $\mathbf{H}$. Moreover, since the background remains static across successive

[1] https://www.ino.ca/en/technologies/video-analytics-dataset
[2] https://www.magnity.com.cn/product/id/76.html

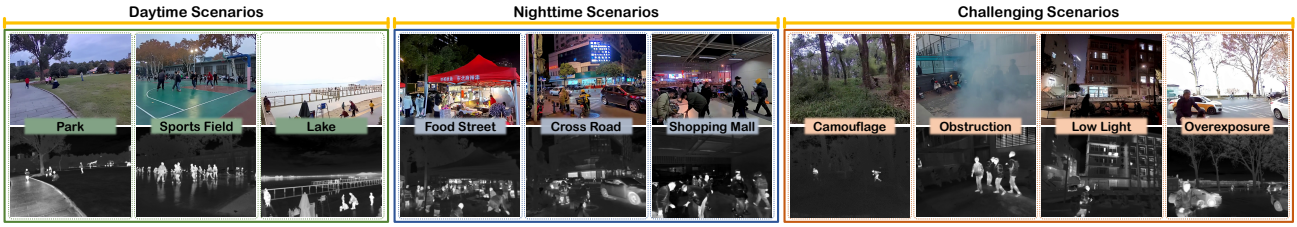

Figure 4. Visualization of various scenarios in our M3SVD dataset.

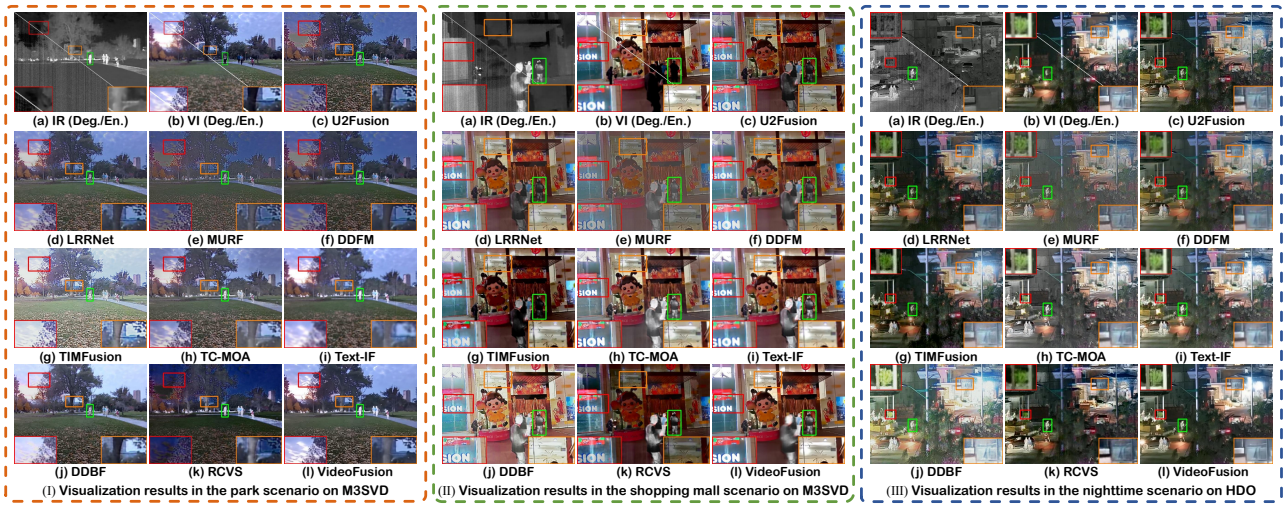

Figure 5. Qualitative comparison results on the M3SVD and HDO datasets under degraded scenarios.

sequences, ensuring a consistent co-visible region, it is reasonable to assume that all frames in a video conform to the same transformation. To enhance the stability of the 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] as well as MINIMA [10], followed by MAGSAC++ [2] to estimate $\mathbf{H}$. Finally, we register each infrared frame to the corresponding downsampled visible frame using the estimated $\mathbf{H}$, resulting in infrared and visible videos with a spatial resolution of 640×480 and a temporal resolution of 30 FPS.

As shown in Fig. 4, 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, covering 100 distinct scenes.

5. Experiments

5.1. Configurations and Implementation Details

The key parameters of our network are set as follows $T = 7$, with N_1, N_2, and N_3 set to 2, 2, and 4, respectively, and C_1, C_2, and C_3 set to 32, 64, and 128, respectively. The hyper-parameters balancing various loss terms 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 with the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized to 1×10^{-4} and gradually reduced to 1×10^{-5} following a cosine annealing schedule. Training is conducted on our M3SVD dataset with 200 videos, where visible and infrared videos are contaminated by blur and stripe noise, respectively. In particular, Gaussian blur with a kernel size of 15 and a standard deviation uniformly sampled from $[0.9, 2.1]$ is applied to degrade visible videos. Moreover, a physics-inspired noise generator [3] is introduced to simulate stripe noise for infrared videos. All experiments are conducted on the pytorch platform using NVIDIA RTX 4090 GPUs and 2.50 GHz Intel(R) Xeon(R) Platinum 8180 CPU.

We compare fusion performance with SOTA image fusion methods, including U2Fusion [41], LRRNet [16], MURF [42], and DDFM [56], DDBF [49], TC-MoA [58], Text-IF [45], TIMFusion [23], as well as video fusion approach, *i.e.* RCVS [40], under degraded and normal scenarios. We also employ SOTA video deblurring (*i.e.*, DSTNet [29]) 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

Table 2. Quantitative comparison results on the M3SVD and HDO datasets under degraded scenarios. Each video in M3SVD and HDO contains 200 and 150 consecutive 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	SF	SD	VIF	SSIM	EN	MI	SF	SD	VIF	SSIM
U2Fusion [41]	6.939	2.502	18.449	35.908	0.441	0.603	6.917	2.068	15.007	49.330	0.402	0.618
LRRNet [16]	6.923	3.136	16.006	38.437	0.454	0.612	6.717	2.279	10.584	43.559	0.341	0.586
MURF [42]	6.216	1.833	14.910	22.843	0.272	0.545	6.485	1.875	14.007	32.493	0.328	0.580
DDFM [56]	6.784	2.669	14.096	32.157	0.450	0.612	6.865	2.084	10.015	43.076	0.379	0.622
DDBF [49]	7.197	2.981	**28.783**	46.584	0.430	0.475	7.260	2.394	**20.394**	50.980	0.411	0.490
TC-MOA [58]	7.131	2.814	15.673	42.622	0.519	0.596	7.303	2.223	14.275	51.011	0.461	0.590
Text-IF [45]	7.195	3.236	9.060	48.829	0.431	0.573	**7.369**	2.473	16.467	53.588	0.460	0.590
TIMFusion [23]	7.099	3.030	18.406	51.080	0.411	0.583	6.688	2.293	11.489	53.169	0.364	0.507
RCVS [40]	6.584	1.989	12.538	37.650	0.473	0.588	7.049	2.015	18.120	53.983	0.419	0.578
VideoFusion	**7.203**	**4.027**	21.552	**52.729**	**0.529**	**0.635**	7.175	**2.927**	18.570	**55.834**	**0.465**	**0.635**

Table 3. Quantitative comparison results on the M3SVD dataset.

Methods	EN	MI	SF	SD	VIF	SSIM
U2Fusion [41]	6.970	2.672	21.091	36.499	0.508	0.617
LRRNet [16]	6.911	3.235	18.982	37.920	0.490	0.621
MURF [42]	6.240	1.924	15.930	22.957	0.312	0.549
DDFM [56]	6.811	2.656	17.960	32.116	0.464	0.596
DDBF [49]	7.229	3.062	35.667	47.834	0.460	0.479
TC-MOA [58]	7.145	2.908	18.619	42.925	0.580	0.605
Text-IF [45]	7.308	3.476	28.104	52.494	0.599	0.602
TIMFusion [23]	7.121	3.098	23.098	51.899	0.435	0.593
RCVS [40]	6.625	2.134	14.641	37.822	0.598	0.609
VideoFusion	7.235	4.212	27.412	53.644	0.608	0.649

dataset with their default configurations.

5.2. Fusion Performance Comparison

Qualitative results under degraded scenarios on the M3SVD and HDO datasets are presented in Fig. 5. As Fig. 5(I) illustrates, U2Fusion, LRRNet, MURF, DDFM, and TIMFusion struggle to preserve prominent targets in infrared videos, with U2Fusion, MURF, and RCVS introducing noticeable artifacts. Additionally, TIMFusion and DDBF adjust scene brightness distribution, causing overexposure in the red-boxed sky. Text-IF can manage image degradations but falters with videos, yielding fusion results with blurring and stripe noise. Similar issues arise in the shopping mall scenario (Fig. 5(II)), where TIMFusion exhibits color distortion, Text-IF retains degradation artifacts, and DDBF overexposes certain areas. In contrast, VideoFusion effectively integrates complementary information and temporal cues, producing clearer and more comprehensive scene representations. For nighttime scenes in HDO (Fig. 5(III)), VideoFusion provides clearer details and highlights significant targets such as pedestrians. Although DDBF brightens the nighttime scene, it introduces overexposure and artifacts along object edges.

We conduct quantitative evaluations on 20 videos from M3SVD and 20 videos with fusion value from HDO, as shown in Tab. 2. VideoFusion achieves the best MI, SSIM, SD, and VIF scores, indicating its ability to transmit multi-source information while maintaining high structural similarity with high-quality source frames. Moreover, our results exhibit superior contrast and better align with human visual perception. Comparable EN values indicate richer information in our fusion results. Notably, DDBF achieves the highest SF by modifying illumination distribution but introduces artifacts, whereas VideoFusion preserves finer details without altering illumination. Additional experiments on conventional M3SVD scenarios (Tab. 3) show VideoFusion excelling in MI, SD, VIF, and SSIM, while slightly trailing Text-IF and DDBF in EN and SF. In summary, both quantitative and qualitative results demonstrate that VideoFusion fully exploits cross-modal complementary information and temporal context to counteract degradation interference and enhance imaging scene characterization.

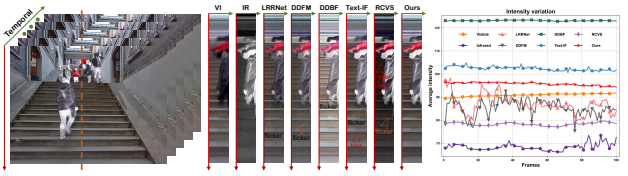

Figure 6. Visual comparisons of the temporal consistency for source and fusion videos. We visualize the pixels of the selected columns (the dotted line) according to [29] and measure the variation of the average brightness across frames. (Please also refer to the videos in the **Supplementary Material**.)

5.3. Temporal Consistency Comparison

While extending image fusion methods to video preserves fusion quality to some extent, it introduces temporal flickering artifacts. Following [29], we show the temporal variation of various videos in Fig. 6, where image-level fusion schemes exhibit noticeable flickering, especially in DDFM

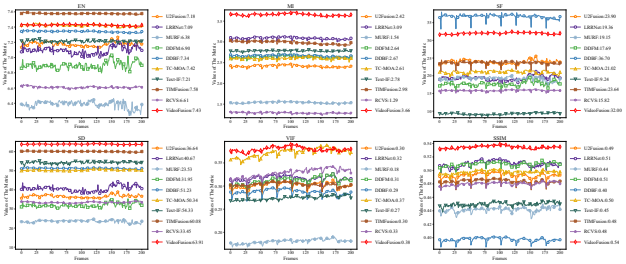

Figure 7. Temporal variation of metrics on consecutive sequences.

Table 4. Comparison of computational efficiency. * denotes methods requiring additional computational complexity of preprocessing algorithms (MDIVDNet and DSTNet) in degraded scenarios.

Methods	MDIVDNet	U2Fusion*	LRRNet*	MURF*	DDFM*	DDBF*
Parm. (M)	10.047	0.659	0.049	0.116	552.660	0.028
Flops (G)	240.15	405.20	14.17	31.49	5220.50	17.10
Time (s)	0.033	0.077	0.011	0.195	3.534	0.033
Methods	**DSTNet**	**Tc-MoA***	**Text-IF**	**TIMFusion***	**RCVS***	**VideoFusion**
Parm. (M)	7.448	340.354	89.014	1.232	0.670	6.743
Flops (G)	224.38	3932.09	1518.88	215.90	27.88	267.78
Time (s)	0.021	0.183	0.055	0.029	0.012	0.070

Figure 8. Visual comparison of object track on M3SVD.

and LRRNet. Although RCVS proposes a unified registration and fusion framework for video streams, its frame-wise fusion strategy inevitably induces flickering. In contrast, our method synthesizes temporally stable fusion results by integrating BiCAM and the variational consistency loss. The stable mean intensity variation further validates the superior temporal consistency property of our approach. Fig. 7 illustrates the temporal curves of various metrics across consecutive sequences, reinforcing this advantage.

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 assessment of various fusion schemes for object tracking on M3SVD, where a pre-trained YOLO v11 [12] 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.

Table 5. Quantitative comparison of ablation studies.

Configs	EN	MI	SF	SD	VIF	SSIM
w/o BiCAM	7.187	3.456	13.730	52.859	0.474	0.603
w/o CmDRM	6.988	3.575	17.942	48.679	0.513	0.615
w/o CMGF	7.082	2.110	31.044	61.714	0.234	0.368
w/o $\mathcal{L}_{grad}$	7.144	3.720	17.438	52.502	0.502	0.618
w/o $\mathcal{L}_{int}$	7.143	3.000	20.678	47.869	0.470	0.633
w/o $\mathcal{L}_{var}$	7.147	3.449	16.209	54.515	0.482	0.602
w/o $\mathcal{L}_{color}$	6.873	2.112	22.456	40.053	0.241	0.459
VideoFusion	7.203	4.027	21.552	52.729	0.529	0.635

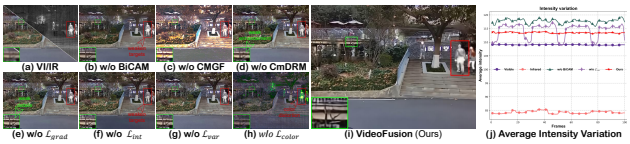

Figure 9. Visual comparison of ablation studies.

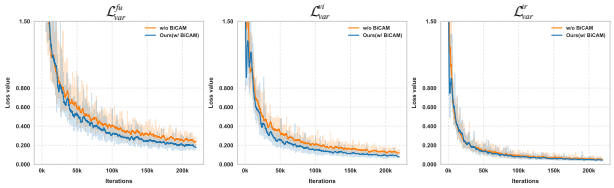

Figure 10. Training processes for BiCAM.

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 $\mathcal{L}_{var}$ 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 SF and 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.

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 *CVPR*, pages 1304–1312, 2020. 6

[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 TIP*, 33:3839–3854, 2024. 3, 6

[4] Tingting Chen, Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T Tan. Dual-rain: Video rain removal using assertive and gentle teachers. In *ECCV*, pages 127–143, 2024. 2, 3

[5] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE TIP*, 32:591–602, 2022. 6

[6] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *ICCV*, pages 12877–12886, 2023. 2, 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[8] Deepak Kumar Jain, Xudong Zhao, Germán González-Almagro, Chenquan Gan, and Ketan Kotecha. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Inf. Fusion*, 95: 401–414, 2023. 1

[9] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, pages 3496–3504, 2021. 5

[10] Xingyu Jiang, Jiangwei Ren, Zizhuo Li, Xin Zhou, Dingkang Liang, and Xiang Bai. Minima: Modality invariant image matching. In *CVPR*, 2025. 6

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

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

[13] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *CVPR*, pages 24966–24976, 2024. 3

[14] Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *ECCV*, pages 433–451, 2024. 3

[15] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 28(5):2614–2623, 2018. 2

[16] 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 TPAMI*, 2023. 2, 6, 7

[17] 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. 2025. 2

[18] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Dynamic scene deblurring by depth guided model. *IEEE TIP*, 29:5273–5288, 2020. 3

[19] 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 *AAAI*, pages 3378–3385, 2024. 2, 3

[20] 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 *CVPR*, pages 5802–5811, 2022. 2, 5

[21] 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 *ICCV*, pages 8115–8124, 2023. 2

[22] 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 *ICCV*, pages 8115–8124, 2023. 5

[23] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. 2024. 2, 6, 7

[24] 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 TIP*, 2023. 2, 3

[25] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for in-

frared and visible image fusion. *Inf. Fusion*, 48:11–26, 2019. 2

[26] 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 TIM*, 70:5009513, 2021. 2

[27] 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 JAS*, 9(7):1200–1217, 2022. 2, 5

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

[29] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *CVPR*, pages 22191–22200, 2023. 2, 3, 6, 7

[30] 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 *ECCV*, pages 421–437. Springer, 2024. 2, 3

[31] Mingyang Song, Yang Zhang, and Tunç O Aydın. Temporformer: Temporally consistent transformer for video denoising. In *ECCV*, pages 481–496. Springer, 2022. 3

[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 JAS*, 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. *Inf. Fusion*, 82: 28–42, 2022. 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. *Inf. Fusion*, 83:79–92, 2022. 2, 4, 5

[35] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Inf. Fusion*, 91:477–493, 2023. 2

[36] 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 *ACM MM*, pages 8546–8555, 2024. 2

[37] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017. 5

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 5

[40] Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. Rcvs: A unified registration and fusion framework for video streams. *IEEE TMM*, 26:11031–11043, 2024. 3, 5, 6, 7

[41] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 44(1):502–518, 2022. 2, 5, 6, 7

[42] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE TPAMI*, 2023. 2, 6, 7

[43] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *AAAI*, pages 6531–6539, 2024. 2, 4

[44] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, pages 3106–3115, 2019. 3

[45] 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 *CVPR*, pages 27026–27035, 2024. 2, 6, 7

[46] 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 *CVPR*, pages 5728–5739, 2022. 5

[47] 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 *AAAI*, pages 12797–12804, 2020. 2

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

[49] 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 *CVPR*, pages 26487–26496, 2024. 2, 6, 7

[50] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *CVPR*, pages 26974–26983, 2024. 2, 5

[51] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE TITS*, 2023. 1

[52] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE TPAMI*, 2023. 1

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

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

[55] 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 *CVPR*, pages 13955–13965, 2023. 2

[56] 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 *ICCV*, pages 8082–8093, 2023. 2, 6, 7

[57] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, pages 2535–2545, 2024. 3

[58] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *CVPR*, pages 7099–7108, 2024. 2, 6, 7