This ICCV 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.

TemCoCo: Temporally Consistent Multi-modal Video Fusion with Visual-Semantic Collaboration

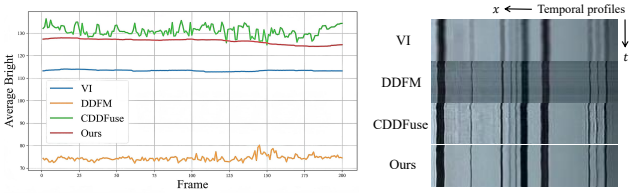
Meiqi Gong Hao Zhang* Xunpeng Yi Linfeng Tang Jiayi Ma*
Electronic Information School, Wuhan University, Wuhan 430072, China
{meiqigong, yixunpeng}@whu.edu.cn, {zhpersonalbox, linfeng0419, jyma2010}@gmail.com

Abstract

Existing multi-modal fusion methods typically apply static frame-based image fusion techniques directly to video fusion tasks, neglecting inherent temporal dependencies and leading to inconsistent results across frames. To address this limitation, we propose the first video fusion framework that explicitly incorporates temporal modeling with visual-semantic collaboration to simultaneously ensure visual fidelity, semantic accuracy, and temporal consistency. First, we introduce a visual-semantic interaction module consisting of a semantic branch and a visual branch, with Dinov2 and VGG19 employed for targeted distillation, allowing simultaneous enhancement of both the visual and semantic representations. Second, we pioneer integrate the video degradation enhancement task into the video fusion pipeline by constructing a temporal cooperative module, which leverages temporal dependencies to facilitate weak information recovery. Third, to ensure temporal consistency, we embed a temporal-enhanced mechanism into the network and devise a temporal loss to guide the optimization process. Finally, we introduce two innovative evaluation metrics tailored for video fusion, aimed at assessing the temporal consistency of the generated fused videos. Extensive experimental results on public video datasets demonstrate the superiority of our method. Our code is released at https://github.com/Meiqi-Gong/TemCoCo.

1. Introduction

The limitations of imaging principles inherently constrain the ability of a single modality to provide a comprehensive representation of a scene [2, 7, 11]. For instance, whether infrared or visible, each modality has its unique strengths and weaknesses [19, 40]. By leveraging the complementary characteristics of these modalities, multi-modal image fusion enables a more thorough and accurate depiction of the scene, addressing the inherent limitations of individual modalities [6, 37]. Owing to these characteristics, fusion images are more beneficial for practical downstream tasks such as autonomous driving [5, 13], security surveillance [14, 17], and others [15, 18, 23].



(a) Average bright variation curves (b) Frame-wise change visualization

Figure 1. Comparison of temporal consistency between two frame-by-frame fusion methods (DDFM and CDDFuse) and our TemCoCo for video fusion. (a): The average brightness variation curves for continuous 200 frames. (b): Visual effects of the same row in continuous 50 frames.

Although significant progress has been made in static frame-based fusion, real-world dynamic environments demand continuous video sequences as the perceptual foundation [3, 38]. In such scenarios, preserving spatial fidelity while ensuring temporal consistency becomes essential [31, 32]. Compared to static frame-based methods, video-based approaches integrate spatiotemporal features, making them more suitable for handling dynamic scene challenges like object motion and environmental changes [12, 35], and thus more applicable to real-world dynamic scenes.

Current research remains predominantly focused on static frame-based image fusion approaches. The rapid evolution of deep learning has revolutionized the paradigm of multi-modal image fusion, marking a significant shift from traditional methods to sophisticated deep learning-based methods [36, 42, 43]. For instance, CrossFuse [9] focuses on mining complementary information between multi-modality image frames by introducing a cross-attention mechanism that captures spatially complementary cues within each frame. SwinFusion [21] introduces a transformer-based framework to enhance fusion by modeling long-range spatial depen-

*Corresponding author.

dencies and capturing both intra-modal and cross-modal interactions within single-frame inputs. Besides, GAN-based and diffusion-based approaches model fusion from distribution perspectives. GANMcC [20] introduces a multi-classification constrained GAN framework, which achieves deep infrared-visible fusion by modeling spatial feature distributions of single-frame inputs. Diff-IF [39] establishes a diffusion paradigm that performs denoising iterations to enable feature interaction within the single-frame spatial domain, progressively coupling two modalities to discover an optimized fusion distribution. Recently, there has been growing awareness of the significance of continuous video fusion [30, 34]. Although several methods have made preliminary attempts to explore video fusion, they fail to modeling temporal dependencies. For example, DFTP applys varying fusion rules across frames to account for temporal fluctuations, rather than leveraging the temporal structure inherent in video data [8]. Consequently, current research on video fusion remains in its infancy, lacking frameworks capable of jointly capturing spatial and temporal dynamics.

Overall, existing methods face challenges in achieving high fusion quality and temporal consistency. These approaches fail to guide and couple visual information with semantic content in a robust manner, thereby compromising semantic accuracy and visual fidelity, which limits subsequent applications, whether human judgment or machine inference [28]. Furthermore, current methods typically treat video fusion as a direct extension of frame-by-frame image fusion [34], neglecting the temporal dependencies between frames, thus exhibiting temporal inconsistency and visible artifacts [3]. To substantiate these claims, Figure 1 presents two state-of-the-art static frame-based fusion methods CDDFuse [44] and DDFM [45], using both quantitative and qualitative evaluations. As shown in Figure 1(a), CDDFuse and DDFM exhibit significant fluctuations in the average brightness variation. In addition, the frame-wise change visualizations in Figure 1(b) further reveal the temporal instability of these methods.

To address the aforementioned challenges, we propose **TemCoCo**, the first temporally-consistent multi-modal video fusion framework that explicitly incorporates temporal modeling and visual-semantic collaboration. TemCoCo is designed to simultaneously capture temporal dynamics and enriched visual-semantic features from degraded source videos, significantly enhancing visual fidelity, semantic accuracy, and temporal consistency. Firstly, we design a novel visual-semantic interaction module comprising a visual branch and a semantic branch. By leveraging the powerful prior knowledge of foundation models (Dinov2 [22] and VGG19), it enhances representational capacity of each branch and ensures meaningful guidance, effectively balancing semantic richness and visual fidelity. To the best of our knowledge, this is the first work to explicitly integrate semantic guidance into video fusion. Secondly, we pioneer the unification of video enhancement and video fusion within a single framework. A temporal cooperative module leverages inter-frame dependencies to recover weak information, while a data augmentation strategy simulates common degradations (*e.g.*, noise and blur) to enhance robustness. This represents the first attempt to jointly address video degradation and video fusion. Thirdly, to enhance temporal consistency, we introduce a temporal-enhanced mechanism coupled with a dedicated temporal consistency loss. The mechanism, comprising 3D convolutions and temporal attention, explicitly models both local and global inter-frame dependencies to capture dynamic scene variations. Meanwhile, the temporal loss enforces temporal alignment by matching frame-wise variations between the fused and source videos. This synergistic combination offers a novel solution for achieving high temporal consistency in results. As illustrated in Figure 1, our proposed TemCoCo yields smoother brightness transitions and higher temporal consistency results than CDDFuse and DDFM, reflecting the effectiveness of our temporal design. Finally, we propose two novel metrics, feaCD and flowD, specifically designed to evaluate temporal coherence in fused videos. feaCD assesses feature-level consistency by measuring the direction of high-level representations across frames, while flowD quantifies image-level smoothness by analyzing optical flow variations.

In summary, our contributions are as follows:

- We propose the first multi-modal video fusion framework that explicitly incorporates both temporal modeling and visual-semantic collaboration, enhancing visual fidelity, semantic accuracy, and temporal consistency.
- We pioneer the unification of video enhancement with video fusion through temporal cooperative modules and targeted guidance, enabling the network to learn robust feature representations in a unified framework.
- We introduce a temporal-enhanced mechanism and a temporal loss, which jointly capture local and global inter-frame dependencies, ensuring smooth and coherent transitions in the fused video.
- We propose two novel temporal metrics to evaluate the temporal consistency between generated and original videos at the feature and image levels, respectively.

2. Method

In this section, we provide a detailed description of our proposed method, TemCoCo, including its overall framework, design details and loss functions.

2.1. Overview

Given a degraded visible video $\tilde{V}_{vis} \in \mathbb{R}^{T\times H\times W\times 3}$ and its corresponding degraded infrared video $\tilde{V}_{ir} \in \mathbb{R}^{T\times H\times W\times 1}$, where T, H and W represent the temporal length and spa-

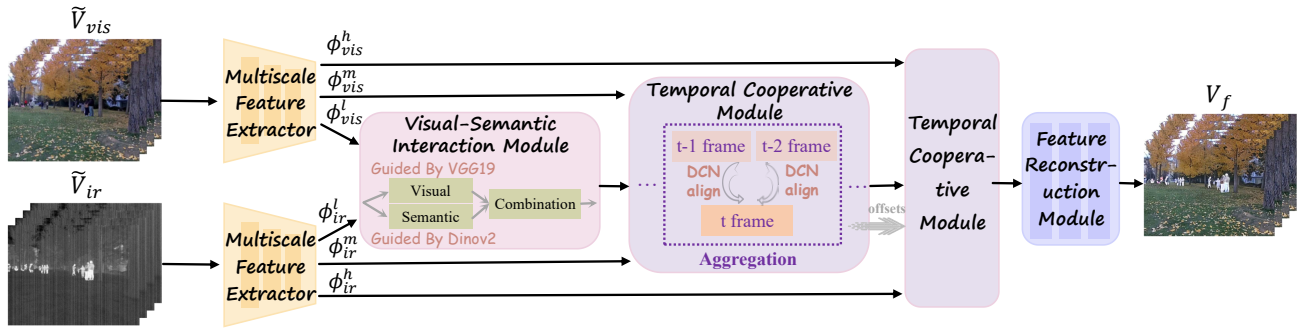
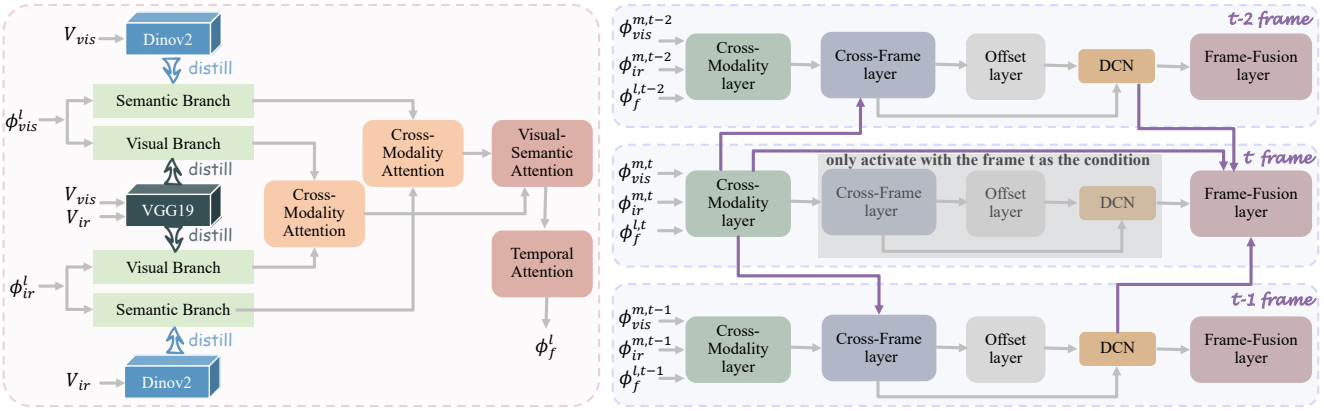


Figure 2. The overall framework of the proposed TemCoCo.



(a) Visual-Semantic Interaction Module

(b) Temporal Cooperative Module

Figure 3. Architectures of (a) Visual-Semantic Interaction Module and (b) Temporal Cooperative Module.

tial dimensions respectively, our objective is to generate a fused video $V_f \in \mathbb{R}^{T\times H\times W\times 3}$ that maintains high fidelity in overall representation, demonstrates superior semantic accuracy, and preserves temporal consistency across frames. We present the overall framework in Figure 2. To achieve spatio-temporal interaction during the fusion process, we activate the temporal-enhanced mechanism to replace the conventional 2D convolutions throughout the network with 3D convolutions. Initially, both $\tilde{V}_{vis}$ and $\tilde{V}_{ir}$ are processed through a Multi-Scale Feature Extractor to obtain hierarchical representations at three distinct scales:

$$\phi^h_{vis}, \phi^m_{vis}, \phi^l_{vis} = E_{vis}(\tilde{V}_{vis}), \tag{1}$$

$$\phi^h_{ir}, \phi^m_{ir}, \phi^l_{ir} = E_{ir}(\tilde{V}_{ir}), \tag{2}$$

where E^{vis} and E^{ir} denote the feature extraction networks for visible and infrared modalities, respectively, and superscripts h, m, and l representing high-, medium-, and low-scale features. The low-scale features $\phi^l_{vis}, \phi^l_{ir}$ are fed into the Visual-Semantic Interaction Module (VSIM) to achieve the separation and aggregation of semantic and visual features. Through targeted guidance from the foundation model, this process enhances both visual quality and semantic accuracy. Subsequently, the medium- and high-scale features are processed by two progressive Temporal Cooperative Modules (TCM), which leverage inter-frame correlations to enhance feature representation. Finally, the enhanced features are reconstructed into the video space through the feature reconstruction module:

$$V_f = RM(\phi^h_f), \tag{3}$$

where RM denotes the feature reconstruction network and ϕ^h_f represents the fused feature. In the following, we will provide detailed descriptions of the VSIM and TCM.

2.2. Visual-Semantic Interaction Module (VSIM)

VSIM is dedicated to extracting enhanced visual-semantic features, with its overall process illustrated in Figure 3(a). First, the module processes low-scale features through dual parallel branches for specialized visual and semantic feature extraction through:

$$\phi^s_{vis}, \phi^s_{ir} = SeB(\phi^l_{vis}, \phi^l_{ir}), \tag{4}$$

$$\phi^v_{vis}, \phi^v_{ir} = ViB(\phi^l_{vis}, \phi^l_{ir}), \tag{5}$$

where SeB and ViB represent the semantic branch network and the visual branch network, respectively. To achieve domain-specific feature enhancement, we strategically integrate VGG19 for visual feature refinement and Dinov2

for semantic understanding, leveraging their powerful priors through knowledge distillation. This integration facilitates the extraction of robust features while preserving domain-specific characteristics. Notably, the input to the foundation models is the enhanced original video, enabling them to assist in recovering features from weak information. Subsequently, the outputs from both modalities through the semantic branch undergo Cross-Modality Attention (CM-Attn) based on SwinTransformer [16] to enhance inter-modal semantic information interaction:

$$\text{CM-Attn}(\phi^x_{vis}, \phi^x_{ir}) = \Big\{ \text{softmax}\Big(\frac{Q_{vis}K_{ir}}{\sqrt{d_k}}\Big)V_{ir}, \\ \text{softmax}\Big(\frac{Q_{ir}K_{vis}}{\sqrt{d_k}}\Big)V_{vis} \Big\}, x \in \{v, s\}, \tag{6}$$

where $Q \in \mathbb{R}^{\frac{HW}{16} \times C}$, $K \in \mathbb{R}^{\frac{HW}{16} \times C}$ and $V \in \mathbb{R}^{\frac{HW}{16} \times C}$ represent the query, key, and value matrices, respectively. The query Q is alternately provided by the visible and infrared modalities, while the key K and value V are supplied by the other modality. The temporal dimension T has been merged into the batch dimension. After that, the cross-modality enhanced features are processed through Semantic-Visual Attention (SV-Attn) to facilitate semantic and visual bidirectional feature interaction. Similar to CM-Attn, the visual and semantic features alternately serve as the query (Q) and key-value pairs (KV). The SV-Attn can be formulated as:

$$\text{SV-Attn}(\phi^v, \phi^s) = \Big\{ \text{softmax}\Big(\frac{Q_v K_s}{\sqrt{d_k}}\Big)V_s, \text{softmax}\Big(\frac{Q_s K_v}{\sqrt{d_k}}\Big)V_v \Big\}. \tag{7}$$

Finally, temporal attention (T-Attn) is applied by flattening features along temporal and channel dimensions to obtain $Q, K, V \in \mathbb{R}^{T \times C}$:

$$\text{T-Attn}(\phi) = \text{softmax}\Big(\frac{Q_t K_t}{\sqrt{d_k}}\Big)V_t. \tag{8}$$

2.3. Temporal Cooperative Module (TCM)

TCM facilitates weak information recovery by leveraging inter-frame correlations, with its detailed architecture presented in Figure 3(b). The diagram specifically demonstrates the processing pipeline for the t-th frame $\phi^{l,t}_f$, while other frames follow an analogous procedure. Notably, the first two frames of the video sequence undergo self-referential enhancement without inter-frame assistance. TCM comprises five fundamental components: cross-modality layer, inter-frame layer, offset layer, DCN [47], and frame fusion layer. The processing pipeline begins with the cross-modality layer receiving medium-scale features ϕ^m_{vis} and ϕ^m_{ir} from the encoder, along with temporally enhanced features ϕ^l_f from VSIM as:

$$\phi_f = CML(\phi^m_{vis}, \phi^m_{ir}, up(\phi^l_f)), \tag{9}$$

where CML denotes the cross-modality layer network and up is the upsampling operation. For the t-th frame ϕ^t_f, we select its two preceding frames ϕ^{t-1}_f and ϕ^{t-2}_f as conditional references to assist in weak information recovery. The inter-frame layer and offset layer learns the spatial offsets o_1 and o_2 from ϕ^{t-1}_f and ϕ^{t-2}_f to ϕ^t_f, enabling frame alignment through DCN, formulated as:

$$\phi^{t-1,t}_f = DCN_1(\phi^{t-1}_f, o_1), \tag{10}$$

$$\phi^{t-2,t}_f = DCN_2(\phi^{t-2}_f, o_2). \tag{11}$$

Finally, the frame fusion layer integrates the features of ϕ^t_f with the aligned features of $\phi^{t-1,t}_f$ and $\phi^{t-2,t}_f$, producing the enhanced feature representation $\phi^{m,t}_f$ with inter-frame assistance, formulated as:

$$\phi^{m,t}_f = FFL(\phi^t_f, \phi^{t-1,t}_f, \phi^{t-2,t}_f), \tag{12}$$

where FFL is the frame fusion layer network. The procedure of TCM at the high scale follows a similar paradigm with the medium scale, differing primarily in the operation of the offset layer. Specifically, the medium-scale processing learns absolute offsets directly, while the high-scale processing learns residual offsets relative to the medium-scale values, thereby enhancing offset precision and computational efficiency.

2.4. Loss Functions

We simulate real-world degradation by applying blur to the original visible videos and adding noise to the original infrared videos, while using original videos as anchors to constrain the network optimization. Initially, we employ the image-level fusion loss commonly used in multi-modal image fusion to constrain the perceptual quality of the fused images, which includes content loss and color loss:

$$\mathcal{L}_{con} = \|\nabla V_f - \max(\nabla V_{vis}, \nabla V_{ir})\|_1 \\ + \|V_f - \max(V_{vis}, V_{ir})\|_1, \tag{13}$$

$$\mathcal{L}_{cor} = \|Cb_f - Cb_{vis}\|_1 + \|Cr_f - Cr_{vis}\|_1. \tag{14}$$

Furthermore, as described in the previous sections, we enhance the feature representation capability of the model through feature distillation from existing foundation models VGG19 and Dinov2. This distillation loss can be formally expressed as:

$$\mathcal{L}_d = \sum_x \|\phi^v_x - VGG19(V_x)\|_2 \\ + \|\phi^s_x - Dinov2(V_x)\|_2, x \in \{vis, ir\}, \tag{15}$$

where ϕ^v_x and ϕ^s_x represent visual features and semantic features from V_x. To better enforce temporal consistency, we further design a temporal consistency loss. It ensures

smooth inter-frame transitions by constraining the variation direction of the fused video to align with that of the original video, defined as:

$$\mathcal{L}_{\mathrm{t}} = \sum_{x}\left(\frac{1}{T-1}\sum_{t=1}^{T-1}\left(1-\cos\left(V_f^t - V_f^{t-1}, V_x^t - V_x^{t-1}\right)\right)\right), \tag{16}$$

where T denotes the number of frames. The final loss function can be concluded as:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda_1\mathcal{L}_{cor} + \lambda_2\mathcal{L}_d + \lambda_3\mathcal{L}_t, \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, which are set to 2, 0.1, and 0.25, respectively.

3. Experiments

In this section, we begin by detailing the implementation specifics and the datasets employed. Next, we introduce the proposed temporal consistency metrics. Subsequently, we compare the video fusion and video segmentation performance of our method with state-of-the-art approaches. Finally, we present ablation studies to validate the effectiveness of each design.

3.1. Datasets and Implementation

Datasets. We employ two infrared and visible video datasets for experiments, named M3SVD[1] and HDO[2]. The M3SVD dataset includes 220 video sequences, with each sequence ranging from 100 to 1000 frames, while the HDO dataset comprises 26 video sequences, each containing between 150 to 200 frames.

Implementation Details. The TemCoCo network is trained for 20 epochs on M3SVD dataset. The learning rate is set to 5×10^{-4}, decaying by a factor of 0.95 every 100 steps. Optimization is performed using the Adam optimizer, with a batch size of 3. The training videos are processed at a resolution of 640×480 with a temporal length of 10 frames, while the testing videos are set to the same resolution with a temporal length of 30 frames. Longer video sequences are constructed through overlapping concatenation to ensure temporal continuity between adjacent video slices. Since our focus is on processing video data with degradation types, such as blur in visible videos and stripe noise in infrared videos, we ensure a fair comparison by pre-processing the original degraded videos for baseline methods that do not consider degradation. Specifically, before performing the fusion, we employ state-of-the-art deblurring methods DSTNet [24] tailored for visible videos and destriping methods MDIVDnet [1] designed for infrared videos, respectively. All experimented are processed on the PyTorch platform using a 2.0-GHz Intel Xeon Gold 6330 CPU and an NVIDIA A800 80GB PCIe GPU.

[1] https://github.com/Linfeng-Tang/M3SVD
[2] https://github.com/xiehousheng/HDO

3.2. Temporal Consistency Metrics

In video fusion, it is essential to consider not only the quality of images but also the temporal consistency across video frames. Existing image fusion metrics fail to address this requirement. Therefore, we propose two novel metrics flowD and feaCD to evaluate inter-frame consistency, considering both the image-level and feature-level perspectives. We provide a comprehensive comparison between our proposed flowD, feaCD and existing video evaluation metrics (tOF [4] and LPIPS [41]) in the supplementary material.

flowD. This metric predicts the next frame of the fused video by leveraging optical flow from the visible video and then compares it with the ground truth. The overall calculation is formulated as follows:

$$\begin{aligned} flowD(V_f, V_{vis}) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\|warp^t - V_f^{t+1}\|_1, \\ warp^t &= warp(V_f^t, flow^t), \end{aligned} \tag{18}$$

where $flow^t = flow(V_{vis}^t, V_{vis}^{t+1})$, and $flow$ is the network of Sea-raft [33], a state-of-the-art optical flow estimation method. Smaller flowD means more temporally consistent with the visible video.

feaCD. This metric evaluates the consistency of change direction between the fused video and the original video at the feature level. Specifically, video features are extracted using ResNet-18, and the cosine similarity between adjacent frames is computed to represent the change direction. The entire process can be formulated as follows:

$$\begin{aligned} \text{feaCD}(V_f, V_{vis}, V_{ir}) &= \frac{1}{T-1}\sum_{x}\sum_{t=1}^{T-1}\|1-\cos_x\|_1, \\ \cos_x &= \cos(F_f^t - F_f^{t-1}, F_x^t - F_x^{t+1}), x \in \{vis, ir\}, \end{aligned} \tag{19}$$

where $F_x^t = ResNet(V_x^t)$ and $F_f^t = ResNet(V_f^t)$. Smaller feaCD indicates more consistent directional changes with the original video.

3.3. Video Fusion

We conduct comparative experiments on video fusion. Due to the scarcity of existing video fusion methods, we select seven state-of-the-art image fusion algorithms and one video fusion algorithm: DATFuse [29], TGFuse [25], DDFM [45], CDDFuse [44], LRRNet [10], DiffIF [39], MMIF-EMMA [46] and RCVS [34]. These algorithms cover CNN-based methods, GAN-based methods, Transformer-based methods, and diffusion-based methods.

We represent qualitative video fusion results on the M3SVD dataset and the HDO dataset in Figure 4 and Figure 5, respectively. We present the enlarged results of the corresponding colored boxes on the right side of each result image. In Figure 4, we first focus on the magnified

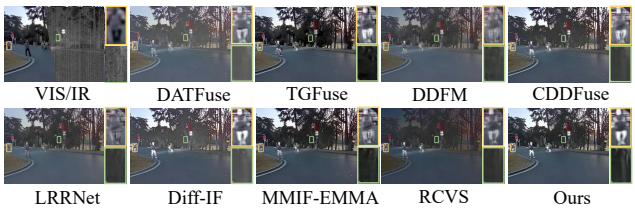


Figure 4. Qualitative video fusion results on the M3SVD dataset.



Figure 5. Qualitative video fusion results on the HDO dataset.

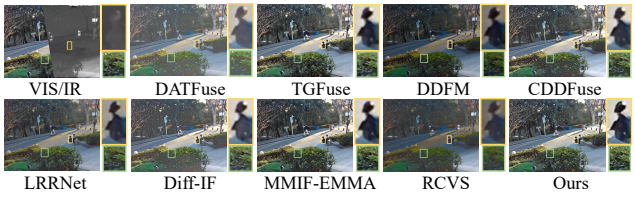


Figure 6. Qualitative video fusion results with real-world degradation.

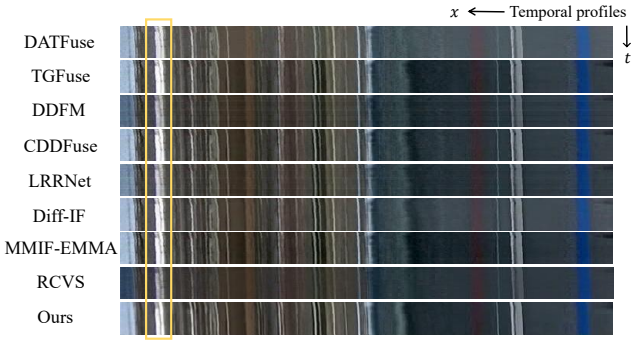


Figure 7. The results of the 150th row from the 10th to the 50th frame extracted along the temporal dimension.

results within the yellow bounding box. Our method demonstrates clearer and more salient targets. Other methods show blurred details and fail to effectively highlight thermal targets. Furthermore, as shown in the green bounding box, only our method successfully preserves the fine details of the tree branches, whereas all other methods lose this critical information. Similarly, in the yellow boxes of Figure 5, only our approach successfully preserves the contour and details of the 'person pouring tea'. In the green boxes, our method also demonstrates the sharpest details, particularly in the wheel of the car. Furthermore, we present experimental results on real-world degraded scenarios, as illustrated in Figure 6. It can be observed that the original visible video suffers from significant blurring. Only our method demonstrates robustness to such degradation, producing clearer fusion results. The enhanced details are particularly evident in the zoomed-in regions.

Table 1. Quantitative video fusion results on the M3SVD dataset. **Bold** is the best.

M3SVD	MI↑	SD↑	flowD↓	feaCD↓
DATFuse	1.9902	32.0269	4.0048	1.8709
TGFuse	2.2379	45.9510	4.3461	1.8338
DDFM	1.6483	26.4525	5.1480	1.9416
CDDFuse	2.1694	45.4648	4.5214	1.9183
LRRNet	2.0207	32.9157	5.0990	1.9417
Diff-IF	2.2491	42.4253	3.7864	1.8346
MMIF-EMMA	2.3140	47.0037	4.3144	1.8598
RCVS	1.1548	29.2992	3.8512	1.8160
Ours	**2.4219**	**50.5400**	**3.2847**	**1.8049**

Table 2. Quantitative video fusion results on the HDO dataset. **Bold** is the best.

HDO	MI↑	SD↑	flowD↓	feaCD↓
DATFuse	1.1115	48.9679	8.0831	1.9791
TGFuse	1.0155	58.4364	8.6420	1.9510
DDFM	1.0265	43.5832	6.9643	1.9714
CDDFuse	1.1396	60.3634	8.3442	1.9460
LRRNet	1.0616	43.3673	7.4598	1.9523
Diff-IF	1.1397	57.4811	8.0450	1.9587
MMIF-EMMA	1.1668	**60.4379**	7.9598	1.9565
RCVS	0.9617	44.2517	7.1108	1.9657
Ours	**1.3373**	59.4020	**5.1147**	**1.7336**

To visualize the advantages of our method in processing video sequences, we present the results of a specific row concatenated along the temporal dimension in Figure 7. The results demonstrate that our method achieves smoother transitions with minimal fluctuations compared to other approaches, highlighting its superior temporal consistency.

Finally, we report quantitative results on the M3SVD and HDO datasets, as shown in Table 1 and Table 2. Four metrics are used: MI (reference-based) and SD (no-reference) for image fusion quality, and flowD and feaCD for temporal consistency. TemCoCo outperforms state-of-the-art methods in both aspects, with notable gains in MI under degraded conditions, highlighting its robustness. Furthermore, our method achieves superior temporal consistency, validating the effectiveness of the temporal designs.

3.4. Video Segmentation

In this part, we present the performance of our proposed TemCoCo compared to other algorithms on the video seg-

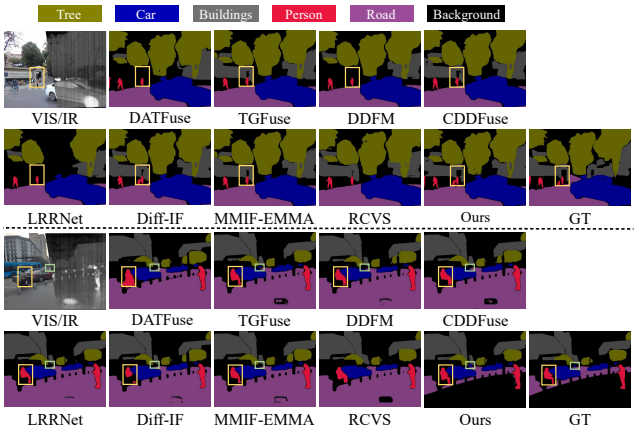


Figure 8. Two examples of video segmentation results.

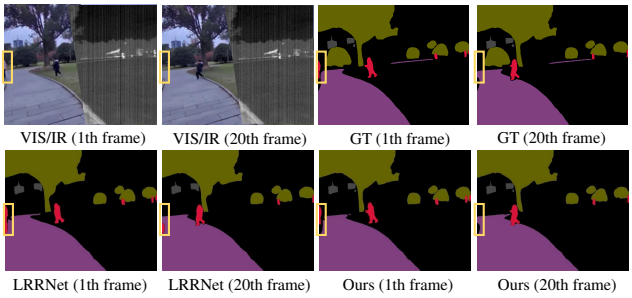


Figure 9. Temporal video segmentation results.

mentation task. Specifically, we employ the Grounded-SAM2 [26, 27] model with text prompts "Road, Person, Tree, Buildings, Car" to perform video segmentation. Due to the lack of a dedicated video segmentation dataset for infrared-visible videos, we manually annotate five video sequences, each comprising 50 frames, with the assistance of SAM2 [26]. The annotations include six classes: road, person, tree, buildings, car, and background.

Figure 8 presents the qualitative segmentation results of two annotated videos alongside their ground-truth labels. In the first example, our method shows superior ability in distinguishing buildings from the background, achieving the highest consistency with the ground truth. In contrast, methods such as DATFuse, DDFM, and LRRNet fail to retain significant parts of the buildings. Meanwhile, TGFuse, CDDFuse, MMIF-EMMA and RCVS are unable to accurately segment the person. Although Diff-IF achieves relatively complete segmentation, it still misses the door part of the building. Similarly, in the second example, only our method and TGFuse successfully segment the car within the green bounding box, and our method is the sole approach that effectively differentiates the road from the background.

Furthermore, in Figure 9, we present an example to highlight the unique challenges of video segmentation compared to image segmentation. Theoretically, only the first frame

Table 3. Quantitative IoU results of video segmentation.

IoU	Road	Person	Tree	Build.	Car	Back.	mIoU
DAT.	0.724	0.733	0.712	0.686	0.754	0.651	0.710
TGF.	0.736	0.779	**0.906**	0.839	0.763	0.724	0.791
DDF.	0.739	0.759	0.876	0.811	0.756	0.716	0.774
CDD.	0.736	0.753	0.735	0.882	0.766	0.806	0.763
LRR.	0.733	0.738	0.829	0.807	0.754	0.672	0.755
Dif.	0.730	0.734	0.824	0.811	0.882	0.765	0.712
MMI.	0.738	0.778	0.726	0.839	0.761	0.691	0.755
RCV.	0.738	0.732	0.868	0.795	0.755	0.706	0.766
Ours	**0.846**	**0.825**	0.901	**0.904**	**0.900**	**0.847**	**0.870**
nodino	0.845	0.710	0.748	0.807	0.695	0.784	0.765

Table 4. Quantitative Acc results of video segmentation.

Acc	Road	Person	Tree	Build.	Car	Back.	mAcc
DAT.	0.925	0.837	0.754	0.805	0.760	0.780	0.810
TGF.	0.957	0.826	**0.932**	0.877	0.771	0.799	0.860
DDF.	**0.959**	0.835	0.918	0.853	0.761	0.810	0.856
CDD.	0.955	0.820	0.773	0.934	0.771	0.806	0.843
LRR.	0.950	0.822	0.849	0.847	0.767	0.811	0.841
Dif.	0.949	0.814	0.863	0.935	0.774	0.798	0.856
MMI.	0.956	0.838	0.767	0.877	0.774	0.798	0.835
RCV.	0.958	0.827	0.903	0.933	0.760	0.775	0.860
Ours	0.957	**0.862**	0.916	**0.942**	**0.914**	**0.920**	**0.919**
nodino	0.942	0.765	0.783	0.850	0.743	0.894	0.830

of video segmentation relies entirely on detection results, while subsequent frames depend on the segmentation outcomes of preceding frames. As shown in the figure, in the 1st and 20th frames, the person on the far left gradually exits the scene. Due to an initial mis-detection by LRRNet, which incorrectly includes part of the step as the person, the 20th frame erroneously continues to classify the step as part of the person. In contrast, our method correctly distinguishes the person in the first frame and maintains accurate segmentation even after the person leaves.

Finally, in Table 3 and Table 4, we report the average segmentation results on the annotated data, including the segmentation accuracy (Acc) and intersection over union (IoU) for each category, as well as the overall mean Acc (mAcc) and mean IoU (mIoU) across all classes. Our method achieves optimal performance in terms of both Acc and IoU across most categories, further demonstrating that our approach effectively integrates semantic information and is applicable to downstream tasks.

3.5. Ablation Studies

We conduct extensive ablation experiments to comprehensively validate the effectiveness of each proposed strat-

Table 5. Quantitative video fusion results of ablation studies.

M3SVD	MI↑	SD↑	flowD↓	feaCD↓
wo_TLoss	2.0142	50.4149	5.7999	1.9374
wo_TAttn	2.1704	50.4297	4.4849	1.8645
no_VSAttn	2.1392	50.2262	4.1371	1.8273
no_DCN	2.2125	50.2244	3.2879	1.8160
no_Dino	2.2591	50.1774	3.2856	1.8075
no_Dino_VGG	2.1612	49.9946	3.2936	1.8124
Conv2d	2.3400	50.2530	3.9472	1.8236
Ours	**2.4219**	**50.5400**	**3.2847**	**1.8049**

egy or module, including the following configurations: **no_TLoss**: Removal of the temporal consistency loss. **no_TAttn**: Removal of the temporal attention mechanism. **no_VSAttn**: Removal of the visual-semantic attention mechanism. **no_DCN**: Replacement of the DCN module with a module that only associates with the current frame. **no_Dino**: Removal of the Dinov2 distillation guidance. **no_Dino_VGG**: Removal of both Dinov2 and VGG19 distillation guidance. **Conv2d**: Replacement of all Conv3d layers with Conv2d layers.

We present the quantitative results for all test videos in Table 5. All ablation variants show a significant drop in image fusion metrics, especially the supervised metric MI, indicating the positive contribution of each module to weak information recovery. However, only a subset of experiments demonstrates a notable impact on temporal consistency, such as no_TLoss, no_TAttn, no_VSAttn, and Conv2d, which aligns well with our design objectives. To further investigate these findings, Figure 10 presents qualitative comparisons of inter-frame smoothness, where clear transition artifacts emerge in the first three configurations. Given the MI degradation across all variants, we also include visual comparisons in Figure 11, where our method produces the clearest textures and most detailed contours, further validating the effectiveness of these components. Finally, we present the impact of removing Dinov2 on segmentation results in the last row of Tables 3 and 4. The results indicate that when the semantic guidance of Dinov2 is eliminated, our method performs comparably to other image-based fusion methods. This is because, although our method excels in temporal consistency, the video segmentation results remain highly dependent on the detection accuracy of the first frame. This finding further validates that the introduction of Dinov2 indeed provides significant semantic information enhancement for the segmentation task, thereby improving the overall segmentation performance.

3.6. Efficiency Comparison

We also give the efficiency comparison results in Table 6. As multi-frame processing shares computation across frames,

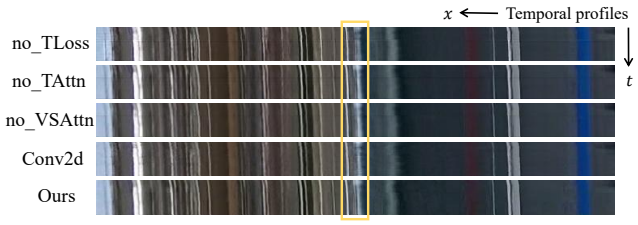


Figure 10. Temporal results of ablation studies.

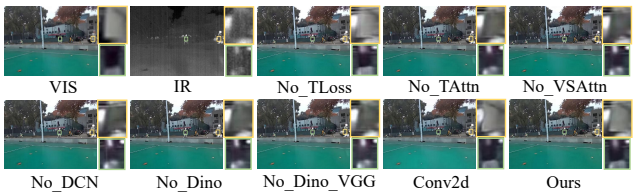


Figure 11. Qualitative video fusion results of ablation studies.

Table 6. Efficiency comparison (FLOPs (G), Time (s)).

Methods	DAT.	TGF.	DDF.	CDD.	LRR.	Dif.	MMI.	RCV.	Ours-2d	Ours
FLOPs	5.6	95.7	5220.5	364.2	14.2	86.9	41.5	**0.05**	442.4	505.6
Time	0.061	0.143	34.494	0.511	0.162	4.233	0.056	**0.009**	0.086	0.109

our TemCoCo remain moderate inference time and acceptable FLOPs. Importantly, our model adopts 3D separable convolutions to effectively capture temporal information with minimal additional overhead, enabling a notable performance gain at relatively low cost. As shown in the table, the computational overhead of using 3D convolutions over 2D counterparts is minor, while the performance improvement shown in Figure 11 is significant, highlighting the favorable trade-off and necessity of our 3D design.

4. Conclusion

This paper presents the first robust multi-modal video fusion framework that explicitly models temporal consistency, while simultaneously addressing the challenges of visual fidelity and semantic accuracy. First, a visual-semantic interaction module, enhanced with Dinov2 and VGG19 distillation, achieves the explicit integration of semantic guidance into video fusion. Second, the temporal cooperative module leverages inter-frame dependencies to recover weak information, pioneering the joint handling of video degradation and fusion in a unified framework. Third, the temporal-enhanced mechanism and the dedicated temporal loss constitute the first explicit design for modeling and enforcing temporal consistency in multi-modal video fusion. Additionally, the two newly introduced metrics provide a more comprehensive and precise quantitative tool for evaluating the temporal consistency. Extensive video experimental results on public datasets demonstrate the superiority. We believe this work opens new avenues for future research in video fusion and related fields.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62276192), and the Fund of National Key Laboratory of Multispectral Information Intelligent Processing Technology (61421132302).

References

[1] 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. 5

[2] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23555–23564, 2023. 1

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 1, 2

[4] Mengyu Chu, You Xie, Laura Leal-Taixé, and Nils Thuerey. Temporally coherent gans for video super-resolution (tecogan). *arXiv preprint arXiv:1811.09393*, 2018. 5

[5] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23 (2):722–739, 2021. 1

[6] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13232–13245, 2023. 1

[7] Meiqi Gong, Hao Zhang, Hebaixu Wang, Jun Chen, Jun Huang, Xin Tian, and Jiayi Ma. Cross-scale domain adaptation with comprehensive information for pansharpening. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 785–793, 2024. 1

[8] 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*, 73:5023520, 2024. 2

[9] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024. 1

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

[11] Xuan Li, Rongfu Chen, Jie Wang, Lei Ma, Li Cheng, and Haiwen Yuan. Dstcfuse: A method based on dual-cycled cross-awareness of structure tensor for semantic segmentation via infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5558–5567, 2024. 1

[12] Hongying Liu, Zhubo Ruan, Peng Zhao, Chao Dong, Fanhua Shang, Yuanyuan Liu, Linlin Yang, and Radu Timofte. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 55(8):5981–6035, 2022. 1

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

[14] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132 (5):1748–1775, 2024. 1

[15] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025. 1

[16] 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/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[17] Ziyi Liu, You Yang, Kejun Wu, Qiong Liu, Xinghua Xu, Xiaoxuan Ma, and Jiang Tang. Asifusion: An adaptive saliency injection-based infrared and visible image fusion network. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–23, 2024. 1

[18] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23 (9):15808–15823, 2022. 1

[19] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1

[20] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:5005014, 2020. 2

[21] 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

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[23] Jia Qi, Tengfei Liang, Wu Liu, Yidong Li, and Yi Jin. A generative-based image fusion strategy for visible-infrared

person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):518–533, 2023. 1

[24] 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, 2024. 5

[25] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023. 5

[26] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 7

[27] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the "edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 7

[28] 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. 2

[29] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. 5

[30] 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

[31] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. Ttnet: Real-time temporal and spatial video analysis of table tennis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 884–885, 2020. 1

[32] Yulin Wang, Yang Yue, Xinhong Xu, Ali Hassani, Victor Kulikov, Nikita Orlov, Shiji Song, Humphrey Shi, and Gao Huang. Adafocusv3: On unified spatial-temporal dynamic video recognition. In *Proceedings of the European Conference on Computer Vision*, pages 226–243, 2022. 1

[33] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 36–54, 2024. 5

[34] 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, 5

[35] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. *IEEE Transactions on Image Processing*, 19(8):2017–2028, 2010. 1

[36] Guoxia Xu, Chunming He, Hao Wang, Hu Zhu, and Weiping Ding. Dm-fusion: Deep model-driven network for heterogeneous image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):10071–10085, 2023. 1

[37] Han Xu, Meiqi Gong, Xin Tian, Jun Huang, and Jiayi Ma. Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Computer Vision and Image Understanding*, 218: 103407, 2022. 1

[38] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 224–242, 2024. 1

[39] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110: 102450, 2024. 2, 5

[40] 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

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[42] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4819–4838, 2021. 1

[43] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023. 1

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

[45] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023. 2, 5

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

[47] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 4