

A Unified Solution to Video Fusion: From Multi-Frame Learning to Benchmarking

Zixiang Zhao¹ Haowen Bai² Bingxin Ke¹ Yukun Cui²
Lilun Deng² Yulun Zhang³ Kai Zhang⁴ Konrad Schindler¹

¹ETH Zürich ²Xi'an Jiaotong University
³Shanghai Jiao Tong University ⁴Nanjing University
zixiang.zhao@ethz.ch

Abstract

The real world is dynamic, yet most image fusion methods process static frames independently, ignoring temporal correlations in videos and leading to flickering and temporal inconsistency. To address this, we propose *Unified Video Fusion (UniVF)*, a novel and unified framework for video fusion that leverages multi-frame learning and optical flow-based feature warping for informative, temporally coherent video fusion. To support its development, we also introduce *Video Fusion Benchmark (VF-Bench)*, the first comprehensive benchmark covering four video fusion tasks: multi-exposure, multi-focus, infrared-visible, and medical fusion. VF-Bench provides high-quality, well-aligned video pairs obtained through synthetic data generation and rigorous curation from existing datasets, with a unified evaluation protocol that jointly assesses the spatial quality and temporal consistency of video fusion. Extensive experiments show that UniVF achieves state-of-the-art results across all tasks on VF-Bench. Project page: vfbench.github.io.

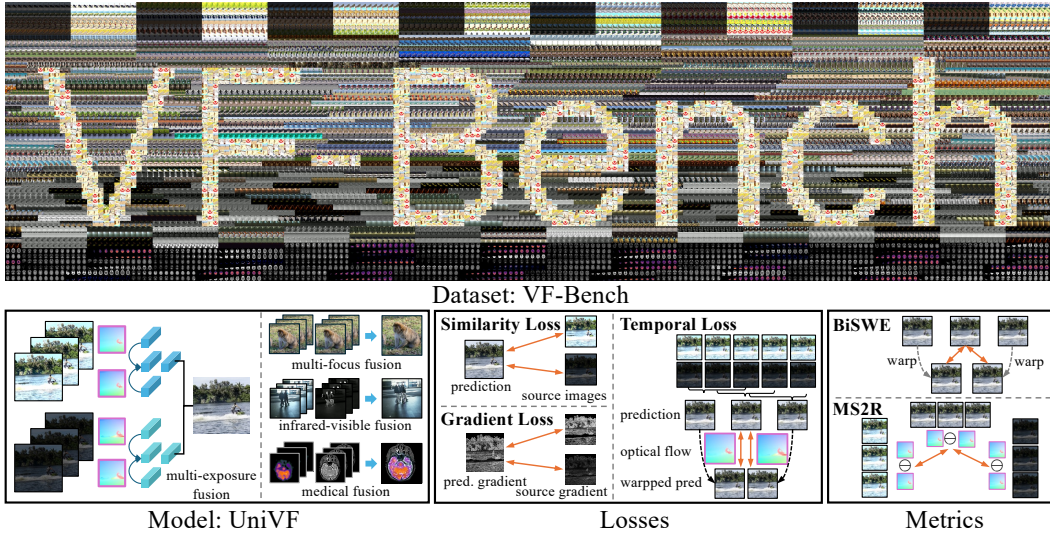


Figure 1: Overview of our main contribution in this paper.

1 Introduction

Image fusion has long been a key research direction in computer vision and image processing. It enables the combination of complementary information from multiple source images into a single,

more informative and perceptually enhanced result [1–6]. A variety of fusion techniques, such as multi-exposure [7, 8], multi-focus [9, 4], infrared-visible [1, 2], and medical image fusion [10, 11], have proven valuable in applications, such as low-light enhancement and exposure correction [12, 13], extended depth-of-field imaging and generation of full-focus scenes [4], target recognition in adverse conditions [14], and improved diagnostic support in clinical imaging [15]. These approaches effectively overcome the limitations of individual sensors or imaging configurations, improving image interpretation by both human observers and machine vision systems. While image fusion has been extensively explored, transitioning to video fusion is a natural next step, as videos provide a continuous and temporally coherent view of dynamic scenes, including object and camera motion, transient events, and contextual variations [16]. With recently advanced imaging hardware and the increasing amount of video data, it has become feasible and necessary to extend image fusion to the temporal domain. The goal is to combine complementary information from multiple input videos into a single, temporally consistent output that offers a more complete representation of the scene.

However, the step from image to video domain introduces several new challenges beyond simply applying image fusion frame-by-frame: (i) *Leveraging temporal information*: Processing frames independently ignores the inherent temporal continuity of videos, leading to flickering and motion discontinuities. Effective video fusion must incorporate information from adjacent frames, not only improving per-frame quality but also ensuring temporal coherence. (ii) *Limited dataset scale*: Compared to paired images, collecting perfectly aligned, temporally synchronized, and diverse video pairs is a lot more challenging and expensive, limiting benchmarking and development for data-driven fusion approaches. (iii) *Lack of evaluation protocols*: Existing evaluation metrics are designed for images, while ignoring consistency along the temporal axis.

To tackle these challenges, we propose a *Unified Video Fusion framework (UniVF)* that explicitly incorporates multi-frame learning to exploit spatial-temporal information, thereby producing informative and temporally consistent fused videos. Specifically, UniVF adopts a Transformer-based [17] encoder-decoder architecture and employs optical flow [18] to warp features from adjacent frames to the current one, effectively capturing temporal dependencies and integrating spatio-temporal relations. A dedicated temporal consistency loss further complements the standard fusion loss based on spatial similarity to suppress flickering and promote temporal continuity across frames.

We then propose a comprehensive *Video Fusion Benchmark (VF-Bench)* that covers four video fusion tasks: *multi-exposure video fusion*, *multi-focus video fusion*, *infrared-visible video fusion*, and *medical video fusion*. For the first two tasks, where paired videos are difficult to acquire directly, we propose novel data generation paradigms: To create multi-exposure data, we utilize 10-bit high dynamic range (HDR) videos, convert the encoded video signals into the linear light domain via the Electro-Optical Transfer Function (EOTF), and perform exposure adjustments to generate diverse exposure pairs; For the multi-focus case, we leverage advances in video depth estimation to simulate the optical focusing process, thereby creating realistic multi-focus video pairs from standard videos. For the latter two tasks, where realistic data synthesis is infeasible, we carefully curate existing datasets by defining objective selection criteria and conducting manual screening to ensure data quality and sufficiently accurate alignment. Moreover, we develop a comprehensive suite of evaluation metrics that cover both the (per-frame) spatial quality and the (frame-to-frame) temporal consistency of a fused video, providing a more holistic evaluation protocol.

Our main contributions can be summarized as follows, with an illustrative overview in Fig. 1:

- We propose a novel *Unified Video Fusion framework, UniVF*, that explicitly incorporates multi-frame learning and cross-frame feature warping to exploit spatial-temporal information, producing informative and temporally consistent videos.
- We construct the first comprehensive *Video Fusion Benchmark, VF-Bench*, by carefully designed data generation strategies and rigorous selection from existing datasets. VF-Bench provides well-aligned, high-quality video pairs across four representative video fusion tasks (multi-exposure, multi-focus, infrared-visible, and medical video fusion).
- To train UniVF on VF-Bench, we introduce a temporal consistency loss alongside the conventional image fusion losses, to suppress flickering and ensure smooth frame transitions in fused videos.
- We establish a comprehensive evaluation protocol for video fusion, integrating both spatial quality and temporal consistency metrics for a thorough assessment.

Experiments on VF-Bench demonstrate that our UniVF achieves state-of-the-art (SOTA) video fusion performance across all four sub-tasks, setting a strong baseline for future research in video fusion.

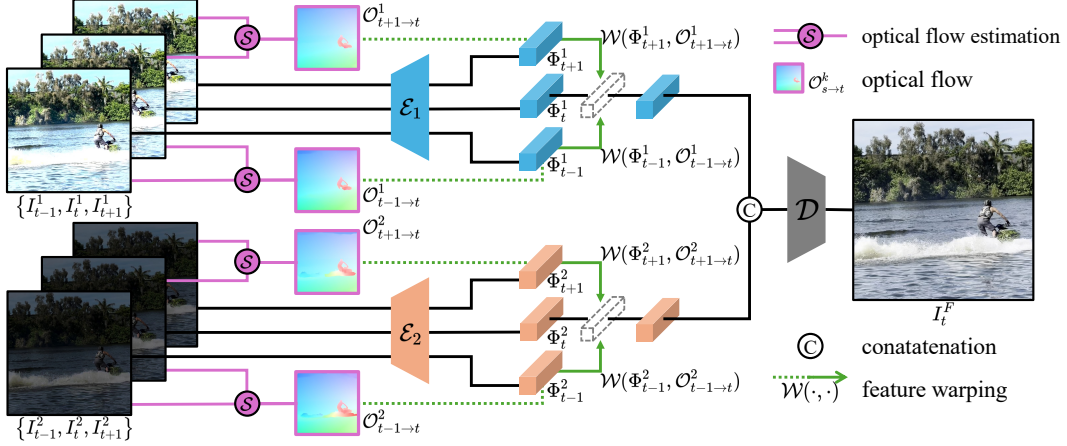


Figure 2: Detailed illustration of our UniVF architecture.

2 Related Work

In the era of deep learning, neural networks are frequently used in image fusion to extract source features, merge features, and reconstruct the fused image [19, 1, 4]. Image fusion algorithms are commonly categorized into two categories: discriminative [20–23] and generative [24, 3]. Discriminative models, utilizing feature extractors such as CNNs [25–30] and Transformers [31, 29], employ model-driven [20–23] or data-driven [32, 1, 33] approaches to obtain features from source images in the image domain, frequency domain, or feature space [34–36]. After information interaction and fusion within the feature space, these models ultimately learn a mapping from the source images to the fused image. On the other hand, generative fusion methods, like GANs [24, 37] and Diffusion [13, 2] models, perform modeling of the latent space manifold to minimize the distributional gap between source and fused images, providing more details in the fused results. Additionally, upstream image registration contributes to robust performance if inputs are misaligned [38–41]. Guidance from downstream tasks [42–44], such as object detection [3, 45, 46] and semantic segmentation [47–50], allows the model to learn more semantically relevant information. Furthermore, unified fusion models leverage inter-task synergy [51–55], while meta-learning can help to better adapt the loss function and the feature extractor [54, 43, 46]. Vision-language models, through their more explicit semantics, offer more flexible guidance [56, 57]. Recently, video fusion, as a further advancement of image fusion, has been demonstrated for infrared and RGB inputs [58, 59]. Moreover, multi-exposure sequences with alternating exposure times enable HDR video reconstruction [60–63]. However, a unified video fusion framework and benchmark are still lacking.

3 UniVF: A Unified Video Fusion Framework

Overview. Given a pair of video sequences $\mathcal{V}_1 = \{I_t^1\}_{t=1}^T$ and $\mathcal{V}_2 = \{I_t^2\}_{t=1}^T$, where T is the total number of frames, the goal of video fusion is to generate a fused video $\mathcal{V}_F = \{I_t^F\}_{t=1}^T$ that integrates complementary information from both inputs. In the following, we introduce our *Unified Video Fusion framework*, **UniVF**, and describe how it utilizes spatial information within each frame and temporal dependencies between adjacent frames to produce temporally coherent fused videos.

3.1 UniVF Details

The proposed UniVF framework is made up of four key components: a feature extractor, an optical flow estimator, a feature warping module, and a feature decoder, which are responsible for extracting frame-wise features, estimating their displacements from frame to frame, aligning them, and reconstructing the fused frames, respectively. An illustration of the architecture is shown in Fig. 2.

Feature Extraction. The goal of this component is to extract domain-specific and spatially rich features from each source video stream. Given a pair $\{\mathcal{V}_1, \mathcal{V}_2\}$, for each time step t , we extract a snippet of three consecutive frames from each source: $\{I_{t-1}^k, I_t^k, I_{t+1}^k\}$ where $k \in \{1, 2\}$. Each video stream has a dedicated encoder $\mathcal{E}_k(\cdot, \cdot, \cdot)$ consisting of several Restormer blocks [17], which is shared

across the three frames of the same source:

$$\Phi_{t-1}^k, \Phi_t^k, \Phi_{t+1}^k = \mathcal{E}_k(I_{t-1}^k, I_t^k, I_{t+1}^k), k \in \{1, 2\}. \quad (1)$$

Optical Flow and Feature Warping. The difference between video fusion and single-frame image fusion lies in the ability to jointly reason over multiple consecutive frames. By exploiting information from preceding and succeeding frames, video fusion can capture dynamics and enhance feature extraction in the current frame. Thus, inspired by [64, 65], we explicitly estimate dense optical flow to align features from adjacent frames to the current time step. Specifically, given two consecutive frames I_s^k and I_t^k ($s \in \{t-1, t+1\}$), SEA-RAFT $\mathcal{S}(\cdot, \cdot)$ [18], a SOTA optical flow estimator, predicts the bidirectional flow $\mathcal{O}_{s \rightarrow t}^k$:

$$\mathcal{O}_{s \rightarrow t}^k = \mathcal{S}(I_s^k, I_t^k), k \in \{1, 2\}, s \in \{t-1, t+1\}. \quad (2)$$

Each optical flow field represents the motion of pixels from one frame to another, where each flow vector indicates the displacement of a pixel to its corresponding location in the neighboring frame. We choose SEA-RAFT [18] for its combination of simplicity, efficiency, and accuracy, which suits our video fusion scenario. Then, to temporally align features, UniVF performs feature warping based on these estimated flows via (differentiable) bilinear sampling. The bidirectional flow fields $\mathcal{O}_{s \rightarrow t}^k$ are used to warp the deep features from adjacent frames to the current time step:

$$\tilde{\Phi}_{s \rightarrow t}^k = \mathcal{W}(\Phi_s^k, \mathcal{O}_{s \rightarrow t}^k), k \in \{1, 2\}, s \in \{t-1, t+1\}, \quad (3)$$

where $\mathcal{W}(\cdot, \mathcal{O})$ denotes warping according to the flow field $\mathcal{O}$. Warped features $\tilde{\Phi}_{s \rightarrow t}^k$ are temporally aligned with the target frame and serve as motion-compensated inputs for subsequent fusion.

Fusion and Reconstruction. The 3×2 feature maps from both sources, warped to a common reference, are concatenated along the channel dimension and fed to the Restormer-based [17] decoder $\mathcal{D}(\cdot)$, which is tasked with modeling long-range dependencies in both the spatial and temporal dimensions:

$$\Phi_t^F = \text{Concat} \left(\Phi_t^1, \Phi_t^2, \tilde{\Phi}_{t-1 \rightarrow t}^1, \tilde{\Phi}_{t+1 \rightarrow t}^1, \tilde{\Phi}_{t-1 \rightarrow t}^2, \tilde{\Phi}_{t+1 \rightarrow t}^2 \right), I_t^F = \mathcal{D}(\Phi_t^F). \quad (4)$$

Finally, the per-frame fusion results I_t^F are reassembled into a fused video sequence.

4 VF-Bench: A Video Fusion Benchmark

Overview. To advance the development and evaluation of video fusion techniques and to promote further research into the topic, we have put together a *Video Fusion Benchmark*, **VF-Bench**, a comprehensive benchmarking suite that includes four different video fusion scenarios: multi-exposure, multi-focus, infrared-visible, and medical video fusion. Examples are shown in Fig. 1. The dataset offers a large collection of paired video sequences with good quality, and precisely aligned to support both model training and testing. In the following we describe our data generation pipeline and the selection criteria. For additional visualizations, as well as further details about data preparation, please refer to Appendices A and B.

Multi-Exposure Video Fusion. To construct multi-exposure video pairs, we propose a novel data processing pipeline with which we synthetically generate different exposure levels from 10-bit HDR source videos by adjusting exposure parameters, see Fig. 3(a). The use of 10-bit HDR videos is advantageous because it preserves a wide dynamic range, ensuring that even after exposure adjustments and potential quality degradation, details are retained that would be lost with 8-bit SDR sources. We start from the YouTube-HDR Dataset [66], a large-scale collection of short-form 10-bit HDR videos sourced from YouTube. From >2000 candidates, we manually curated 500 scenes with an average of 150 frames, choosing those with rich visual content, vivid colors, and free from watermarks or video effects. These were further divided into 450 for training and 50 for testing.

Since exposure depends approximately linearly on scene radiance, it is essential to perform exposure adjustment in the linear light domain. In this way one accurately simulates radiometric changes and avoids distortions introduced by non-linear gamma-encoding. Therefore, we first convert the encoded video signals with the Electro-Optical Transfer Function (EOTF) and transform the video into a linear color space. Exposure adjustments are then applied in this linear domain by ± 3 EV (exposure value), simulating over-exposed and under-exposed conditions. To produce 8-bit videos,

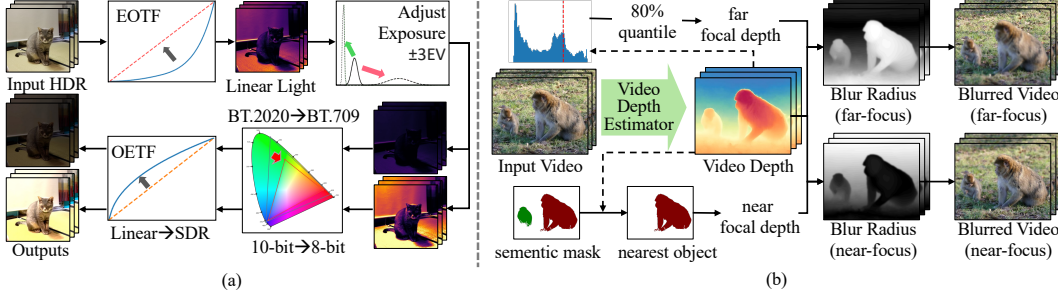


Figure 3: The proposed data generation paradigms for (a) multi-exposure video pair and (b) multi-focus video pair for our VF-Bench.

we then perform gamut mapping from the 10-bit BT.2020 color space to the 8-bit BT.709 color space. The adjusted videos are mapped to 8-bit SDR video pairs using the BT.709 standard Gamma Opto-Electronic Transfer Function (OETF) [67], to obtain paired over- and under-exposed video sequences. The described process closely resembles real-world multi-exposure video capture, where a single scene is recorded at different exposure settings, while ensuring consistent color mapping and precise spatial-temporal alignment across exposure levels. More details about this process are provided in Appendix A.1.

Multi-Focus Video Fusion. Existing multi-focus image fusion datasets primarily rely on light field cameras [68, 69], manually labeled focal masks [70, 71] or blur simulation based on foreground-background semantic segmentation masks [72]. The former two approaches are expensive and difficult to scale up, whereas the last one does not follow the physical process: focal planes and associated circles of confusion (CoC), by definition, depend on continuous scene depth [73] rather than on semantic labels.

Therefore, we propose to utilize depth maps to construct a multi-focus video dataset, by estimating the per-pixel blur radius, see Fig. 3(b). We use the DAVIS dataset [74] as video source for our experiments. Specifically, 150 videos are split into 120 training scenes and 30 test scenes, with an average length of 70 frames. We run a single-view video depth estimator [75] on them to obtain dense (inverse) depth. Given a focal depth, the CoC for pixel i can be calculated as:

$$CoC^i = Af|(D_f - D^i)/(D_f - f)|/D^i \approx d_f|1 - d^i/d_f|\sigma, \quad (5)$$

with A the aperture, f the focal length, D_f the focal depth, and D^i the depth of scene pixel i . To account for unknown camera metadata, the CoC is approximated by the estimated normalized inverse depth d^i , the given normalized focal depth d_f , and a constant blur strength factor σ . Further details on the derivation of Eq. (5) can be found in Appendix A.2.

We select two normalized focal depth values, d_f^{far} and d_f^{near} , from the first frame of each video, representing the background and foreground focus, respectively. The background focal depth is the 20th percentile of the inverse depth values. To find the foreground focus depth, we compute an average depth value for each segmented object according to the DAVIS masks [74] and select the closest of them. Finally, a Gaussian blur is applied to each pixel with the kernel size proportional to the calculated CoC, thereby generating paired multi-focus video sequences.

Infrared-Visible Video Fusion. Unlike the previous two tasks, paired infrared-visible video datasets can be obtained from RGBT tracking benchmarks, whereas they are difficult to realistically simulate. To construct our video fusion dataset, we start from the VT MOT [76] dataset and implement a three-stage filtering process to ensure data quality, accurate alignment, and diversity of content. First, infrared video frames are evaluated using three objective metrics: Image Entropy, Global Contrast, and Dark Area Proportion. Frames that exhibit low entropy, insufficient contrast or excessive dark regions are considered low-quality and discarded, thus removing uninformative scenes. Second, for the RGB modality, we adopt Retinex theory to decompose each frame into illumination and reflectance components [77]. RGB frames with high illumination values, indicating sufficient ambient lighting and limited need for infrared data, are excluded, thus retaining only video pairs where the infrared and visible channels provide complementary information. Finally, we perform frame-wise fusion of the remaining video pairs using SOTA image fusion algorithms like CDDFuse [1] and EMMA [78], followed by manual inspection to validate the alignment and eliminate pairs affected by mis-registration or ghosting artifacts. Further details of the complete selection criteria can be found

in Appendix A.3. Through this process, we curate a total of 90 video scenes with, on average, 300 frames. These are randomly split into 75 training scenes and 15 testing scenes.

Medical Video Fusion. For medical video fusion, we rely on the Harvard Medical dataset [79] as a source, treating consecutive slices of MRI and corresponding CT, PET or SPECT volumes as video sequences. A filtering strategy similar to the one used for infrared-visible fusion is adopted to ensure data quality. Specifically, frames with large invalid regions or poor visual quality are removed, preserving only meaningful, interpretable sequences with rich visual details. As a result, we curate a total of 57 scenes with 27 frames on average, which are divided into 49 for training and 8 for testing.

5 Experiments

We now evaluate our UniVF on VF-Bench for all four fusion scenarios: multi-exposure fusion (MEF), multi-focus fusion (MFF), infrared-visible fusion (IVF), and medical video fusion (MVF). We first describe the experimental setup, with a particular focus on the newly proposed temporal consistency term in the training loss, as well as the temporal consistency evaluation metrics that we add to the single-frame evaluation protocol. Then, we discuss the results, which highlight that our approach already constitutes a strong baseline. Finally, we conduct ablation studies to validate our design choices. Further experimental results are shown in Appendix D due to space limitations.

5.1 Setup

Loss Function. To jointly optimize spatial fidelity and temporal consistency, we adopt a compound training loss with three terms:

$$\mathcal{L} = \mathcal{L}_{\text{spatial}} + \alpha_1 \mathcal{L}_{\text{grad}} + \alpha_2 \mathcal{L}_{\text{temp}}, \quad (6)$$

where $\{\alpha_1, \alpha_2\}$ are weight parameters, set to $\{10, 2\}$, $\{1, 0.5\}$, $\{5, 2\}$ and $\{1, 1\}$ for MEF, MFF, IVF and MVF tasks respectively, such that the terms have comparable magnitudes. The three losses are (i) *Spatial similarity*: the spatial loss $\mathcal{L}_{\text{spatial}}$ measures per-pixel reconstruction error. Specifically, for IVF and MVF tasks, following [1], $\mathcal{L}_{\text{spatial}} = \frac{1}{HW} \|I_t^F - \max(I_t^1, I_t^2)\|_1$. For MEF tasks, following [56], we set $\mathcal{L}_{\text{spatial}} = \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{MEF-SSIM}}$, where $\mathcal{L}_{\text{int}} = \frac{1}{HW} \|I_t^F - \text{mean}(I_t^1, I_t^2)\|_1$, and $\mathcal{L}_{\text{MEF-SSIM}}$ borrows from [80]. For MFF task, we set $\mathcal{L}_{\text{spatial}} = \mathcal{L}_{\text{int}} = \frac{1}{HW} \|I_t^F - \text{mean}(I_t^1, I_t^2)\|_1$. (ii) *Gradient preservation*: to preserve image structures and edges, following [56], we introduce a dedicated gradient loss $\mathcal{L}_{\text{grad}} = \frac{1}{HW} \|\nabla I_t^F - \max(|\nabla I_t^1|, |\nabla I_t^2|)\|_1$. ∇ denotes the Sobel gradient operator. (iii) *Temporal consistency*: To suppress flickering and ensure smooth transitions across frames, we introduce a temporal consistency loss that explicitly enforces frame-to-frame consistency, by penalizing misalignments between adjacent frames:

$$\begin{aligned} \mathcal{L}_{\text{temp}} = & \mathbb{E}_{p \in M_{\text{prev}}^t} [\|I_t^F(p) - \mathcal{W}(I_{t-1}^F, \mathcal{O}_{t-1 \rightarrow t}^F)(p)\|_1] \\ & + \mathbb{E}_{p \in M_{\text{next}}^t} [\|I_t^F(p) - \mathcal{W}(I_{t+1}^F, \mathcal{O}_{t+1 \rightarrow t}^F)(p)\|_1], \end{aligned} \quad (7)$$

where I_t^F is the fused video sequence from Eq. (4), and $\mathcal{W}(\cdot, \mathcal{O})$ denotes warping with the optical flow field $\mathcal{O}$. M_{prev}^t and M_{next}^t are validity masks at time step t that indicate regions with reliable flow estimates, as detailed below. This loss term implements the strong temporal continuity of videos with reasonable frame rates, by enforcing consistency between the current fused frame and its two warped neighbors, so as to reduce flickering and abrupt changes.

Validity masks. To improve the robustness of the temporal consistency loss term $\mathcal{L}_{\text{temp}}$ in Eq. (7) and avoid unreliable gradients, we introduce validity masks $\{M_{\text{prev}}^t, M_{\text{next}}^t\}$ to identify well-aligned, non-occluded regions between adjacent frames. Each mask is derived via a forward-backward flow consistency check: given the forward flow $\mathcal{O}_{t \rightarrow t+1}$ and backward flow $\mathcal{O}_{t+1 \rightarrow t}$, the latter is first warped onto the coordinate space of frame t as $\hat{\mathcal{O}}_{t+1 \rightarrow t}(p) = \mathcal{W}(\mathcal{O}_{t+1 \rightarrow t}, \mathcal{O}_{t \rightarrow t+1}(p))$. Since the backward flow is defined in the pixel space of frame $t+1$ while the forward flow originates from frame t , warping enables both flows to be compared in the same coordinate space. The consistency error is computed as $\Delta(p) = \|\mathcal{O}_{t \rightarrow t+1}(p) + \hat{\mathcal{O}}_{t+1 \rightarrow t}(p)\|_2$, and the binary mask is defined as $M(p) = 1$ if $\Delta(p) < \epsilon$, otherwise 0. ϵ is a predefined threshold (set to 1.0 in our implementation). Intuitively, this verifies whether a pixel can move forward and then return along the estimated flow paths with minimal deviation. Large inconsistencies typically indicate occlusions, motion boundaries, or flow



Figure 4: Previous, current, and next frames with their corresponding validity masks M_{prev}^t and M_{next}^t . Black regions denote invalid or unreliable areas, corresponding to poorly aligned or occluded pixels that are excluded from the temporal consistency computation.

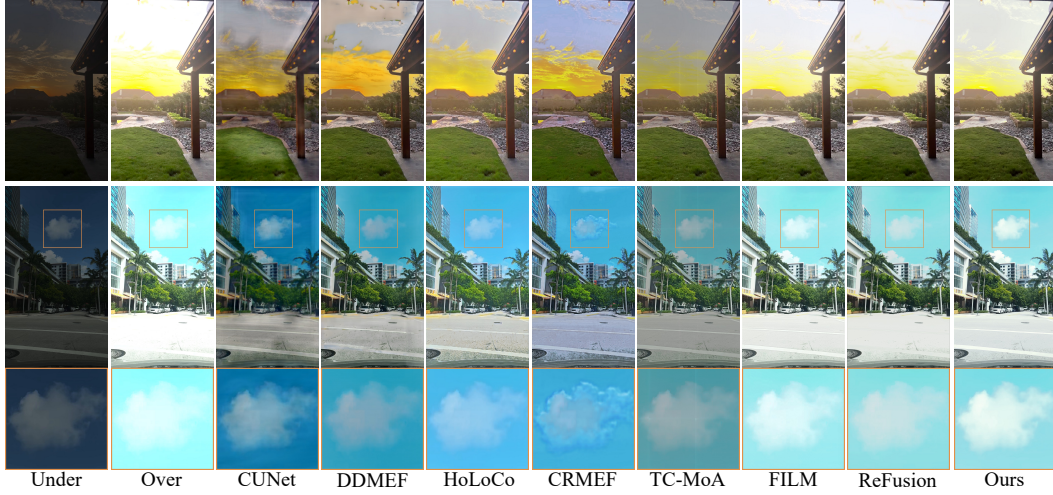


Figure 5: Qualitative comparison of fusion outcomes for multi-exposure video fusion.

ambiguities. By restricting the temporal consistency loss to these valid regions, UniVF effectively reduces flickering and enforces smooth temporal transitions. Visual examples of $\{M_{prev}^t, M_{next}^t\}$ are shown in Fig. 4.

Training details. We ran our experiments on a machine equipped with a single NVIDIA GeForce RTX 4090 GPU. The loss is minimized with Adam, starting with a learning rate of 10^{-4} that decays exponentially to 1% of its initial value over the course of 20k iterations. Training uses a batch size of 32, with gradient accumulation. As our network architecture, we adopt Restormer blocks [17] in both the encoder $\mathcal{E}_k(\cdot)$ and decoder $\mathcal{D}(\cdot)$ components. Each block contains 8 attention heads and has a feature dimension of 32. Both the encoders and decoder are configured with 4 stacked blocks.

Metrics. (i) *Spatial domain evaluation metrics:* We adopt four widely used quantitative metrics: VIF (Visual Information Fidelity), SSIM (Structural Similarity Index), MI (Mutual Information), and $Q^{AB/F}$. These indicators measure perceptual fidelity, structural similarity, mutual information content, and edge preservation, respectively. In all cases, a higher value means better fusion of the complementary information from the two sources. For further details, see [81]. (ii) *Temporal domain consistency metrics:* To assess the temporal consistency and motion smoothness of fused videos, we proposed two complementary evaluation metrics.

Bi-Directional Self-Warping Error (BiSWE): As a reference-free metric, BiSWE is designed to quantify frame-to-frame *temporal alignment errors* within a video sequence. Given a video clip $\{I_{t-1}, I_t, I_{t+1}\}$, we compute the optical flow fields $\mathcal{O}_{s \rightarrow t} = \mathcal{S}(I_s, I_t)$, $s \in \{t-1, t+1\}$ with SEA-RAFT $\mathcal{S}(\cdot, \cdot)$ [18]. Validity masks $\{M_{prev}^t, M_{next}^t\}$ are applied to exclude unreliable regions based on forward-backward consistency. The BiSWE value is computed as

$$\begin{aligned} \text{BiSWE} = & \mathbb{E}_{p \in M_{prev}^t} [|I_t(p) - \mathcal{W}(I_{t-1}, \mathcal{O}_{t-1 \rightarrow t})(p)|_1] \\ & + \mathbb{E}_{p \in M_{next}^t} [|I_t(p) - \mathcal{W}(I_{t+1}, \mathcal{O}_{t+1 \rightarrow t})(p)|_1], \end{aligned} \quad (8)$$

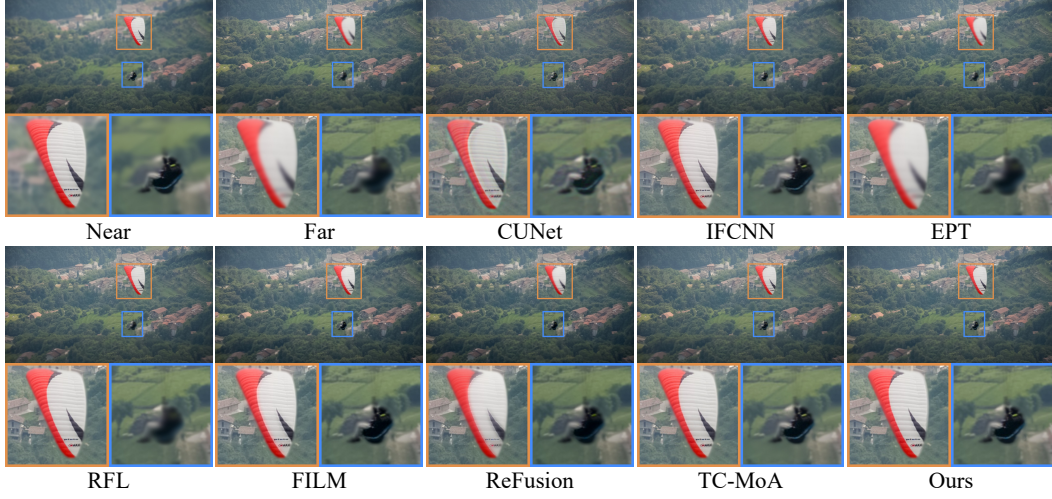


Figure 6: Qualitative comparison of fusion outcomes for multi-focus video fusion.

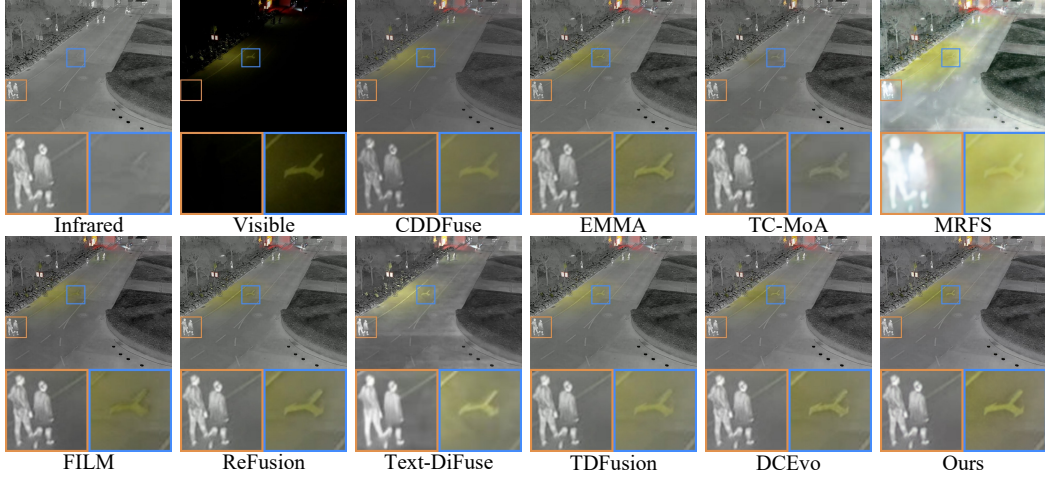


Figure 7: Qualitative comparison of fusion outcomes for infrared-visible video fusion.

where $\mathcal{W}(\cdot, \mathcal{O})$ is the warping function guided by the flow $\mathcal{O}$, and $\mathbb{E}$ denotes averaging over valid pixels. Lower BiSWE indicates improved temporal alignment.

Motion Smoothness with Dual Reference Videos (MS2R): To assess the naturalness and consistency of motion transitions, MS2R evaluates the coherence of *flow changes* in the fused video and two reference sequences. The flow change is defined as the difference between consecutive flows within a clip. Intuitively, it compares the accelerations of objects projected onto the image plane. The MS2R score is defined as

$$\text{MS2R} = \mathbb{E}_p [|\Delta \mathcal{O}^F(p) - \Delta \mathcal{O}^{R_1}(p)|_1] + \mathbb{E}_p [|\Delta \mathcal{O}^F(p) - \Delta \mathcal{O}^{R_2}(p)|_1], \quad (9)$$

where $\Delta \mathcal{O}^F = \mathcal{O}_{1 \rightarrow 2}^F - \mathcal{O}_{0 \rightarrow 1}^F$, $\mathcal{O}_{1 \rightarrow 2}^F$ and $\mathcal{O}_{0 \rightarrow 1}^F$ are also obtained from $\mathcal{S}(\cdot, \cdot)$ [18]. $\Delta \mathcal{O}^{R_1}, \Delta \mathcal{O}^{R_2}$ are computed similarly from the two reference sequences. A lower MS2R indicates smoother and more natural motion trajectories in the fusion video.

5.2 Video Fusion Experiments

Multi-Exposure Video Fusion. We ran experiments on the MEF branch of VF-Bench. Tested methods include CUNet [9], DDMEF [82], HoLoCo [83], CRMEF [84], TC-MoA [55], FILM [56] and ReFusion [54]. As illustrated in Tab. 1 and Fig. 5, UniVF consistently achieves superior

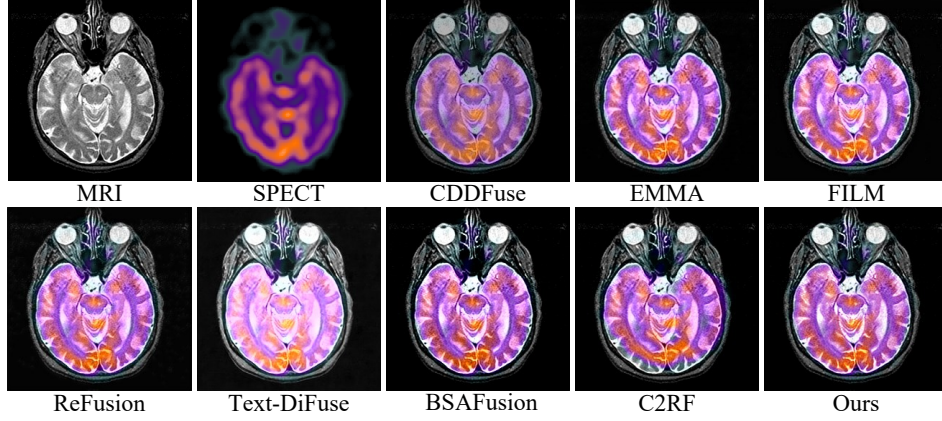


Figure 8: Qualitative comparison of fusion outcomes for medical video fusion.

Table 1: Quantitative evaluation results for the MEF and MFF task. The red and blue highlights indicate the highest and second-highest scores.

VF-Bench Multi-Exposure Video Fusion Branch							VF-Bench Multi-Focus Video Fusion Branch						
	VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓		VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓
CUNet	0.58	0.67	2.26	0.39	7.12	0.42	CUNet	0.56	0.88	3.74	0.53	6.79	1.47
DDMEF	0.72	0.94	2.71	0.65	10.06	1.04	IFCNN	0.69	0.89	4.93	0.70	6.25	1.38
HoLoCo	0.39	0.79	2.30	0.25	10.52	0.55	RFL	0.78	0.90	6.29	0.78	5.96	1.11
CRMEF	0.64	0.95	2.61	0.61	8.68	0.42	EPT	0.77	0.90	6.31	0.78	5.97	1.14
TC-MoA	0.76	0.98	2.94	0.71	7.78	0.34	TC-MoA	0.77	0.90	5.46	0.76	5.99	1.13
FILM	0.78	0.98	4.39	0.71	8.27	0.34	FILM	0.76	0.89	5.02	0.75	6.32	1.28
ReFus	0.75	0.97	3.89	0.70	7.95	0.33	ReFus	0.73	0.90	4.95	0.73	5.80	1.28
Ours	0.82	0.99	4.45	0.72	6.40	0.33	Ours	0.79	0.90	6.32	0.79	5.95	1.08

quantitative and qualitative performance, effectively balancing dynamic range expansion, contrast enhancement, and image quality preservation across multiple exposure levels.

Multi-Focus Video Fusion. For the experiments on the MFF branch of VF-Bench, the tested methods include CUNet [9], IFCNN [85], RFL [86], EPT [87], TC-MoA [55], FILM [56], and ReFusion [54]. As shown in Fig. 6 and Tab. 1, our UniVF baseline again achieves the best performance both qualitatively and quantitatively, accurately identifying focused regions and producing sharp fusion results free of artifacts, across both the foreground and background.

Notably, both training and testing for the MEF and MFF branches were conducted on the 2K-resolution version of the dataset. Considering potential computational resource constraints, we also report results on a low-resolution version in the Appendix C for reference.

Infrared-Visible Video Fusion. We conducted experiments on the IVF branch of VF-Bench, with CDDFuse [88], EMMA [14], TC-MoA [55], MRFS [50], FILM [56], ReFusion [54], Text-DiFuse [57], TDFusion [43] and DCEvo [42]. The results once more show superior performance of UniVF, in terms of both visual quality and quantitative metrics. As illustrated in Fig. 7, the method faithfully preserves critical thermal and structural details, enhances object visibility and reduces noise in low-light conditions. Quantitative comparisons in Tab. 2 further confirm that UniVF consistently has an edge in most metrics, highlighting its robustness across diverse scenes and object types.

Medical Video Fusion. On the MVF branch of VF-Bench, we evaluate a range of methods including CDDFuse [88], EMMA [14], FILM [56], ReFusion [54], Text-DiFuse [57], BSAFusion [89] and C2RF [38]. As can be seen in Fig. 8 and Tab. 3, UniVF once more effectively preserves fine-grained textures from MRI images, while simultaneously enhancing and maintaining the salient intensities of the CT, PET or SPECT modality. The fused results exhibit clear anatomical details and clean tissue boundaries from the MRI source, along with distinct color distributions originating from the SPECT images to support clinical diagnosis.

Ablation Studies. To explore the contribution of each key component within UniVF, we conducted ablation studies on the IVF task, with results summarized in Tab. 4. In Exp. I, we removed the feature warping module, *i.e.*, multi-frame features are directly concatenated along the channel dimension and fed into the decoder, without optical flow correction. In Exp. II, both the feature warping module

Table 2: Quantitative evaluation for the IVF task.

VF-Bench Infrared-Visible Video Fusion Branch						
	VIF $\uparrow$	SSIM $\uparrow$	MI $\uparrow$	Qabf $\uparrow$	BiSWE $\downarrow$	MS2R $\downarrow$
CDDF	0.37	0.64	2.41	0.54	5.12	0.37
EMMA	0.37	0.63	2.01	0.58	4.79	0.37
TC-MoA	0.37	0.64	2.05	0.60	4.68	0.38
MRFS	0.27	0.55	1.48	0.34	6.09	0.38
FILM	0.40	0.63	2.05	0.64	4.78	0.37
ReFus	0.42	0.64	2.27	0.67	4.64	0.36
Text-D	0.30	0.60	1.64	0.39	10.63	0.40
TD Fusion	0.45	0.64	2.34	0.67	4.35	0.36
DCEvo	0.43	0.64	2.44	0.66	4.57	0.37
Ours	0.44	0.64	2.47	0.68	3.94	0.35

Table 3: Quantitative evaluation for the MVF task.

VF-Bench Medical Video Fusion Branch						
	VIF $\uparrow$	SSIM $\uparrow$	MI $\uparrow$	Qabf $\uparrow$	BiSWE $\downarrow$	MS2R $\downarrow$
CDDF	0.29	0.76	1.80	0.59	26.33	1.34
EMMA	0.29	0.68	1.73	0.60	30.00	1.98
FILM	0.33	0.36	1.83	0.67	32.04	1.59
ReFus	0.31	0.32	1.74	0.67	32.85	1.74
Text-D	0.24	0.21	1.58	0.52	34.09	1.96
BSAF	0.28	0.63	1.69	0.58	34.73	1.66
C2RF	0.30	0.73	1.75	0.59	32.67	2.06
Ours	0.35	0.76	2.00	0.68	29.61	1.30

Table 4: Ablation experiments results, with red representing the best values.

Descriptions	Configurations			Metrics					
	feature warping	multi-inputs	$\mathcal{L}_{temp}$	VIF $\uparrow$	SSIM $\uparrow$	MI $\uparrow$	Qabf $\uparrow$	BiSWE $\downarrow$	MS2R $\downarrow$
Exp. I: w/o feature warping		✓	✓	0.40	0.63	2.44	0.66	4.18	0.36
Exp. II: w/o warping & multi-inputs			✓	0.38	0.61	2.07	0.64	4.46	0.37
Exp. III: w/o $\mathcal{L}_{temp}$	✓	✓		0.42	0.65	2.38	0.65	5.79	0.39
UniVF (Ours)	✓	✓	✓	0.44	0.64	2.47	0.68	3.94	0.35

and multi-frame inputs were removed, reverting the model to a conventional frame-by-frame fusion scheme. To ensure a fair comparison, we increased the number of Restormer blocks to maintain an equivalent total parameter count. In Exp. III, while retaining the original multi-frame fusion structure with feature warping, we switched off the temporal consistency loss $\mathcal{L}_{temp}$ during training.

Taken together, the ablation experiments demonstrate the necessity of each component in our scheme. Specifically, multi-frame feature warping enhances temporal coherence and overall fusion quality, while the temporal consistency loss further ensures smooth transitions across consecutive frames. Their combination yields superior video fusion compared to simplified or frame-wise baselines.

6 Conclusion

We have presented **UniVF**, a unified framework for video fusion that explicitly leverages multi-frame learning and optical flow-based feature warping to exploit both spatial and temporal information. To ensure temporal coherence of the fused results, we introduce a custom temporal consistency loss that suppresses flickering and enforces smooth frame transitions. Additionally, we have introduced a comprehensive **VF-Bench**, to our knowledge the first benchmark for video fusion. It covers four representative tasks with carefully constructed, paired video datasets and a holistic evaluation protocol, including dedicated temporal consistency metrics. Extensive experiments demonstrate that UniVF, with its straightforward design, achieves SOTA performance across all four tasks. We hope that our benchmark, together with the strong baseline of our fusion framework, encourages further research into video fusion and lays a solid foundation for it.

Acknowledgments

This work was supported by Huawei Technologies Oy (Finland), and by a SwissAI Compute Grant.

References

- [1] Zixiang Zhao, Haowen Bai, Jianshe 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5906–5916, 2023.
- [2] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jianshe 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 *Int. Conf. Comput. Vis.*, pages 8082–8093, 2023.
- [3] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5792–5801, 2022.

- [4] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4819–4838, 2021.
- [5] Zhengxue Wang, Zhiqiang Yan, Jinshan Pan, Guangwei Gao, Kai Zhang, and Jian Yang. Dornet: A degradation oriented and regularized network for blind depth super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15813–15822, 2025.
- [6] Zeyu Xiao and Xinchao Wang. Event-based video super-resolution via state space models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12564–12574, 2025.
- [7] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Trans. Image Process.*, 29:2808–2819, 2020.
- [8] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: A structural patch decomposition approach. *IEEE Trans. Image Process.*, 26(5):2519–2532, 2017.
- [9] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3333–3348, 2020.
- [10] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Inf. Fusion*, 19:4–19, 2014.
- [11] Han Xu and Jiayi Ma. Emfusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion*, 76:177–186, 2021.
- [12] Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, and Jinwei Gu. Polarfree: Polarization-based reflection-free imaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10890–10899, 2025.
- [13] Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16111–16121, June 2025.
- [14] Zixiang Zhao, Haowen Bai, Jianshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25912–25921. IEEE, 2024.
- [15] Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K. Kasabov. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Trans. Neural Networks Learn. Syst.*, pages 1–14, 2023.
- [16] Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, et al. USTC-TD: A test dataset and benchmark for image and video coding in 2020s. *IEEE Trans. Multimedia*, 2025.
- [17] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5718–5729, 2022.
- [18] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: simple, efficient, accurate RAFT for optical flow. In *Eur. Conf. Comput. Vis.*, volume 15065, pages 36–54. Springer, 2024.
- [19] Julius Erbach, Dominik Narnhofer, Andreas Dombos, Bernt Schiele, Jan Eric Lenssen, and Konrad Schindler. Solving inverse problems with flair. *arXiv preprint arXiv:2506.02680*, 2025.
- [20] Shuang Xu, Jianshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1366–1375, 2021.
- [21] Zixiang Zhao, Shuang Xu, Jianshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2022.
- [22] Hui Li, Tianyang Xu, Xiaojun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11040–11052, 2023.
- [23] Zixiang Zhao, Jianshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5687–5697, 2022.

- [24] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion*, 48:11–26, 2019.
- [25] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Int. Conf. Comput. Vis.*, pages 4724–4732, 2017.
- [26] Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Inf. Fusion*, 79:248–262, 2022.
- [27] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Trans. Image Process.*, 31:1325–1339, 2022.
- [28] Odysseas Bouzos, Ioannis Andreadis, and Nikolaos Mitianoudis. A convolutional neural network-based conditional random field model for structured multi-focus image fusion robust to noise. *IEEE Trans. Image Process.*, 32:2915–2930, 2023.
- [29] Yuanshen Guan, Ruikang Xu, Mingde Yao, Lizhi Wang, and Zhiwei Xiong. Mutual-guided dynamic network for image fusion. In *ACM Int. Conf. Multimedia*, pages 1779–1788, 2023.
- [30] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Trans. Image Process.*, 30:3098–3112, 2021.
- [31] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *AAAI*, pages 2126–2134, 2022.
- [32] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jianshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020.
- [33] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.*, 28(5):2614–2623, 2018.
- [34] Haoyu Ma, Qingmin Liao, Juncheng Zhang, Shaojun Liu, and Jing-Hao Xue. An α -matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Trans. Image Process.*, 29:8668–8679, 2020.
- [35] Zeyu Wang, Xiongfei Li, Haoran Duan, and Xiaoli Zhang. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Trans. Image Process.*, 31:4527–4542, 2022.
- [36] Zeyu Wang, Xiongfei Li, Libo Zhao, Haoran Duan, Shidong Wang, Hao Liu, and Xiaoli Zhang. When multi-focus image fusion networks meet traditional edge-preservation technology. *Int. J. Comput. Vis.*, pages 1–24, 2023.
- [37] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping (Steven) Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.*, 29:4980–4995, 2020.
- [38] Linfeng Tang, Qinglong Yan, Leyuan Fang Xinyu Xiang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *Int. J. Comput. Vis.*, 2025.
- [39] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022.
- [40] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM Int. Conf. Multimedia*, pages 3783–3791, 2022.
- [41] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19647–19656, 2022.
- [42] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2226–2235, 2025.
- [43] Haowen Bai, Jianshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7457–7468, 2025.

- [44] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Baisong Jiang, Lilun Deng, Yukun Cui, Shuang Xu, and Chunxia Zhang. Deep unfolding multi-modal image fusion network via attribution analysis. *IEEE Trans. Circuit Syst. Video Technol.*, 35(4):3498–3511, 2025.
- [45] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *ACM Int. Conf. Multimedia*, pages 4003–4011, 2022.
- [46] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13955–13965. IEEE, 2023.
- [47] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.
- [48] 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.
- [49] 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 *Int. Conf. Comput. Vis.*, pages 8115–8124, 2023.
- [50] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 26974–26983, 2024.
- [51] Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito Ali, Muhammad Awais, and Josef Kittler. One model for ALL: low-level task interaction is a key to task-agnostic image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 28102–28112, 2025.
- [52] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2022.
- [53] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Eur. Conf. Comput. Vis.*, pages 719–735, 2022.
- [54] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *Int. J. Comput. Vis.*, 133(5):2547–2567, 2025.
- [55] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7099–7108. IEEE, 2024.
- [56] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *Int. Conf. Mach. Learn.*, 2024.
- [57] Hao Zhang, Lei Cao, and Jiayi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. In *Adv. Neural Inform. Process. Syst.*, 2024.
- [58] Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025.
- [59] Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. Rcvs: A unified registration and fusion framework for video streams. *IEEE Trans. Multimedia*, 2024.
- [60] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24851–24860, 2024.
- [61] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K. Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Int. Conf. Comput. Vis.*, pages 2502–2511, 2021.
- [62] Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world hdr video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2879–2888, 2024.
- [63] Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *Int. Conf. Comput. Vis.*, pages 12760–12769, 2023.

- [64] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5972–5981, 2022.
- [65] Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *IEEE Trans. Image Process.*, 33:2171–2182, 2024.
- [66] Yilin Wang, Joong Gon Yim, Neil Birkbeck, and Balu Adsumilli. Youtube SFV+HDR quality dataset. In *IEEE Int. Conf. Image Process.*, pages 96–102. IEEE, 2024.
- [67] BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange. <https://www.itu.int/rec/R-REC-BT.2100>, 2018. Accessed: 2021-02-02.
- [68] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Inf. Fusion*, 25:72–84, 2015.
- [69] Juncheng Zhang, Qingmin Liao, Shaojun Liu, Haoyu Ma, Wenming Yang, and Jing-Hao Xue. Real-mff: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138:370–377, 2020.
- [70] Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jianshe Zhang. Mffw: A new dataset for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*, 2020.
- [71] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Inf. Fusion*, 66:40–53, 2021.
- [72] Xiaopeng Guo, Rencan Nie, Jinde Cao, Dongming Zhou, Liye Mei, and Kangjian He. Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network. *IEEE Trans. Multimedia*, 21(8):1982–1996, 2019.
- [73] R. Fernando. *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Addison-Wesley Professional, 2004.
- [74] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- [75] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.
- [76] Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, Chengjie Gu, and Zhixiang Huang. Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognit.*, 161:111330, 2025.
- [77] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22252–22261. IEEE, 2023.
- [78] Zixiang Zhao, Haowen Bai, Jianshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25912–25921. IEEE, 2024.
- [79] BKeith A. Johnson and J. Alex Becker. Harvard medical website. <http://www.med.harvard.edu/AANLIB/home.html>.
- [80] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.*, 24(11):3345–3356, 2015.
- [81] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion*, 45:153–178, 2019.
- [82] Xiao Tan, Huaian Chen, Rui Zhang, Qihan Wang, Yan Kan, Jinjin Zheng, Yi Jin, and Enhong Chen. Deep multi-exposure image fusion for dynamic scenes. *IEEE Trans. Image Process.*, 32:5310–5325, 2023.
- [83] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Inf. Fusion*, 95:237–249, 2023.

- [84] Zhu Liu, Jinyuan Liu, Guanyao Wu, Zihang Chen, Xin Fan, and Risheng Liu. Searching a compact architecture for robust multi-exposure image fusion. *IEEE Trans. Circuit Syst. Video Technol.*, 34(7):6224–6237, 2024.
- [85] 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.
- [86] Zeyu Wang, Xiongfei Li, Haoran Duan, and Xiaoli Zhang. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Trans. Image Process.*, 31:4527–4542, 2022.
- [87] Zeyu Wang, Xiongfei Li, Libo Zhao, Haoran Duan, Shidong Wang, Hao Liu, and Xiaoli Zhang. When multi-focus image fusion networks meet traditional edge-preservation technology. *Int. J. Comput. Vis.*, 131(10):2529–2552, 2023.
- [88] Zixiang Zhao, Haowen Bai, Jianshe 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5906–5916. IEEE, 2023.
- [89] Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *AAAI*, pages 4725–4733, 2025.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: We have clearly stated the contributions and scope in the abstract and introduction. The claims made are consistent with both the theoretical discussion and experimental results presented in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have discussed the limitations of our work in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There are no new theoretical results in our paper, which focuses mainly on method design, benchmark establishment, and empirical evaluation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have detailed the experimental settings, including datasets, model configurations, and evaluation metrics, in the main paper and Appendix to ensure reproducibility. Additionally, our code has been made publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided instructions for reproduction details in the main paper and Appendix. The dataset, code and pretrained models are publicly available to ensure faithful reproduction of the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All necessary details, including data splits, hyperparameters, and optimizer settings, are provided in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars, but provide extensive experimental results across multiple benchmark datasets to demonstrate consistency and robustness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have details this in the main paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research does not involve sensitive data, human subjects, or potential ethical concerns, and conforms fully with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper is a basic study on video fusion and does not directly involve societal impacts to the best of our knowledge.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We have not released any data or models that may have a risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We have cited all original papers that produced the code packages and datasets used in our work, and respected their licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We have publicly released all models, datasets, and pretrained weights, along with detailed documentation and example scripts to ensure reproducibility and ease of use.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects, so no IRB approval is required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We used LLM models only for writing and formatting assistance, which did not affect the core methodology or research outcomes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Further Details of Data Preparation in VF-Bench

A.1 Explanation of Key Terms in Multi-Exposure Video Fusion

- **BT.709**: A widely adopted International Telecommunication Union Radiocommunication Sector (ITU-R) standard for high-definition television (HDTV) video, specifying parameters such as color primaries, transfer characteristics (OETF), and color space for 8-bit SDR (Standard Dynamic Range) video.
- **BT.2020**: An ITU-R recommendation defining parameters for ultra-high-definition television (UHDTV), including a wider color gamut, higher bit-depth (typically 10-bit or 12-bit), and enhanced color reproduction capabilities compared to BT.709. It is used for HDR (High Dynamic Range) video content.
- **Electro-Optical Transfer Function (EOTF)**: A mathematical function that defines how digital signal values are converted into visible light by a display. It transforms non-linear, gamma-encoded video signals into a linear light domain, accurately representing scene radiance. BT.2020 (HLG encoding format¹) EOTF is defined as:

$$L = \begin{cases} \frac{V^2}{3} & 0 \leq V \leq 0.5 \\ \frac{\exp(\frac{V-0.5599}{0.1788}) + 0.2847}{12} & 0.5 < V \leq 1 \end{cases} \quad (10)$$

where V is the normalized video signal value and L is the corresponding linear luminance.

- **Opto-Electronic Transfer Function (OETF)**: The inverse of EOTF, this function defines how light captured by a camera sensor is converted into digital video signal values. It typically applies a gamma curve to map linear scene radiance into a non-linear encoding space suitable for storage and broadcast. The OETF for BT.709 is specified as:

$$V = \begin{cases} 4.5L & 0 \leq L < 0.018 \\ 1.099L^{0.45} - 0.099 & 0.018 \leq L \leq 1 \end{cases} \quad (11)$$

where L is the normalized linear light level and V is the video signal value.

- **Linear Color Space**: A color space where the numerical values of pixel intensities are directly proportional to the physical light intensity in the real world. In this domain, exposure adjustments and radiometric operations can be performed accurately, as opposed to gamma-encoded, non-linear spaces where such operations would introduce distortions.

A.2 Circle of Confusion Derivation and Approximation (Eq. (5)) in Multi-Focus Video Fusion

The blur level in optical imaging systems is characterized by the size of the Circle of Confusion (CoC), which determines the degree of defocus blur for each pixel. A larger CoC corresponds to stronger blur. Based on the thin lens equation:

$$\frac{1}{v} + \frac{1}{D} = \frac{1}{f}, \quad (12)$$

where v is the image distance, D is the object distance, and f is the focal length of the lens. The image distance for an object at depth D^i is given by:

$$v^i = \frac{fD^i}{D^i - f}. \quad (13)$$

Assuming a focus distance D_f , the corresponding image distance is:

$$v_f = \frac{fD_f}{D_f - f}. \quad (14)$$

As illustrated in Fig. 9, the CoC at pixel i can then be computed as:

$$\text{CoC}^i = A \left| \frac{v^i - v_f}{v^i} \right|, \quad (15)$$

¹Hybrid Log-Gamma (HLG) is a widely used high dynamic range (HDR) encoding format, employed in the YouTube-HDR dataset [66].

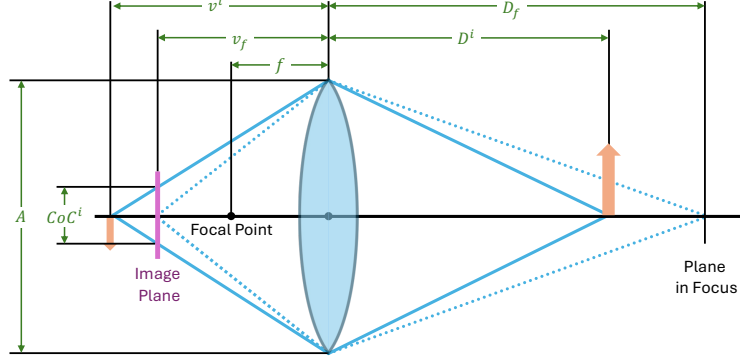


Figure 9: Illustration of the optical geometry for Circle of Confusion.

where A is the aperture diameter.

Substituting the expressions for v^i and v_f , and simplifying, we obtain:

$$\text{CoC}^i = Af \left| \frac{D^i - D_f}{D^i(D_f - f)} \right|. \quad (16)$$

To facilitate practical implementation without requiring precise camera metadata, we approximate the CoC based on estimated normalized inverse depth. Let $d^i = 1/D^i$ and $d_f = 1/D_f$. Assuming $f \ll D^i, D_f$ (a valid assumption in most video capture scenarios), and noting that the focus distance D_f remains fixed within a specific video, we approximate Eq. (16) as:

$$\text{CoC}^i \propto |D^i - D_f|/D^i = \left| \frac{1}{d^i} - \frac{1}{d_f} \right| \cdot d^i. \quad (17)$$

Further simplifying yields:

$$\text{CoC}^i \propto \left| 1 - \frac{d^i}{d_f} \right|. \quad (18)$$

By introducing a global blur strength factor σ to account for unknown camera parameters, the CoC for pixel i is approximated by:

$$\text{CoC}^i \approx d_f \left| 1 - \frac{d^i}{d_f} \right| \sigma. \quad (19)$$

This approximation is justified as it preserves the relative CoC values across pixels, which is critical for simulating realistic defocus blur patterns in the absence of explicit optical parameters. Furthermore, since inverse depth typically correlates linearly with perceived defocus in monocular video sequences, this approximation remains perceptually valid. The scaling factor σ absorbs the unknown optical constants and ensures consistent blur strength across frames. We take $\sigma = 0.025$, which is approximated by using a common camera setup.

Finally, for a frame with the longer edge length l pixels, we calculate the Gaussian blur kernel size kernel^i for each pixel from the calculated CoC values by:

$$\text{kernel}^i = \text{CoC}^i \cdot l. \quad (20)$$

A.3 Selection Criteria for Infrared-Visible Video Fusion

A.3.1 Objective Assessment for Infrared Frames

To ensure the quality of infrared-visible video fusion, an objective assessment is performed on infrared frames prior to fusion. Three quantitative metrics — *Image Entropy*, *Global Contrast*, and *Dark Area Proportion* — are used to evaluate each frame. Frames that do not meet predefined thresholds are

discarded to exclude low-quality or uninformative content. In the following, we present more details of the metrics, computation methods, and thresholds.

Image Entropy. Image Entropy quantifies the information content and textural complexity of a grayscale image. Higher entropy indicates richer pixel intensity distribution, while lower values indicate less informative content. It is computed as:

$$H = - \sum_{i=0}^{255} p(i) \log_2 p(i), \quad (21)$$

where $p(i)$ is the normalized histogram value of intensity i . The entropy of each infrared frame is calculated from its normalized histogram via Eq. (21). Frames with $H > 6$ are retained while those below this threshold are excluded due to insufficient information.

Global Contrast. Global Contrast is assessed by the standard deviation of pixel intensities, reflecting the overall contrast distribution in the image. A higher standard deviation indicates stronger contrast and clearer object boundaries, which is essential to highlight thermal patterns. It is computed as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}, \quad (22)$$

where x_i is the intensity of pixel i , μ the mean intensity, and N the total number of pixels. Each frame's contrast is evaluated from its grayscale intensities. Frames with $\sigma > 30$ are retained; those below are discarded due to insufficient structural and thermal contrast.

Dark Area Proportion. Frames containing excessive dark regions often reflect poor capture conditions or insufficient thermal signals. To quantify this, we compute the *Dark Area Proportion* D , which represents the proportion of pixels whose intensity falls below a predefined threshold $T = 10$:

$$D = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(I_i \leq T), \quad (23)$$

where N is the total number of pixels in the frame and $\mathbb{I}(\cdot)$ is the indicator function:

$$\mathbb{I}(I_i \leq T) = \begin{cases} 1, & \text{if } I_i \leq T \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

Frames with a dark ratio under 5% are retained while those exceeding this value are discarded for lacking sufficient thermal content and visibility.

Overall Quality Scoring. After filtering based on each of the three individual metrics, we further perform a comprehensive selection of scenes using a weighted normalized score that combines the three metrics:

$$\text{Score} = w_1 \cdot \frac{H}{H_{\max}} + w_2 \cdot \frac{\sigma}{\sigma_{\max}} + w_3 \cdot (1 - D). \quad (25)$$

$H_{\max}$ and $\sigma_{\max}$ denote the maximum values of H and σ over the entire dataset, respectively. Weights w_1 , w_2 , and w_3 satisfying $w_1 + w_2 + w_3 = 1$. Equal weights ($w_1 = w_2 = w_3 = 1/3$) are used here. This composite score offers an intuitive measure for evaluating overall frame quality. Based on this score, we discard the bottom 10% of scenes.

A.3.2 Exclusion of High-Illumination RGB Frames

Retinex theory decouples a natural image (I) into an illumination component (L) and a reflectance component (R), mathematically expressed as:

$$I = L * R. \quad (26)$$

Here, L represents the intensity of illumination incident upon the scene, which varies with lighting conditions. Conversely, R signifies the intrinsic reflectance properties of the scene, which remain invariant to changes in illumination. This Retinex decomposition allows for the effective extraction of the scene's illumination information, facilitating subsequent filtering processes. In our work, we leverage the model proposed in [77] to extract the scene's illumination component. Following this, frames with illumination levels ranked in the top 25% based on $\text{mean}(L)$ values are excluded.

Table 5: Quantitative evaluation results for the low-resolution MEF (540p) and MFF (480p) task. The red and blue highlights indicate the highest and second-highest scores.

VF-Bench Multi-Exposure Fusion Branch (540p)							VF-Bench Multi-Focus Fusion Branch (480p)						
	VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓		VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓
CUNet	0.50	0.85	1.85	0.39	7.55	0.20	CUNet	0.53	0.86	3.52	0.68	10.23	0.42
DDMEF	0.71	0.95	2.96	0.66	8.99	0.71	IFCNN	0.68	0.87	4.85	0.73	9.37	0.38
HoLoCo	0.50	0.86	2.56	0.42	8.22	0.19	RFL	0.77	0.90	6.31	0.78	8.46	0.28
CRMEF	0.62	0.94	2.60	0.63	8.72	0.19	EPT	0.76	0.90	6.33	0.78	8.50	0.29
TC-MoA	0.74	0.99	2.93	0.72	7.82	0.16	TC-MoA	0.75	0.90	5.27	0.77	8.39	0.28
FILM	0.77	0.99	4.35	0.72	8.28	0.17	FILM	0.75	0.89	5.06	0.78	8.61	0.33
ReFus	0.74	0.97	3.81	0.72	7.63	0.16	ReFus	0.73	0.90	4.93	0.77	8.00	0.32
Ours	0.79	0.99	4.38	0.73	6.96	0.16	Ours	0.77	0.90	6.34	0.79	8.29	0.27

B Visualizations of the Four Branches in VF-Bench

We present visualizations of some part of the video pairs from the four branches in VF-Bench as follows:

- Dataset visualizations for *Multi-exposure Video Fusion* branch in VF-Bench are shown in Fig. 10.
- Dataset visualizations for *Multi-focus Video Fusion* branch in VF-Bench are shown in Fig. 11.
- Dataset visualizations for *Infrared-Visible Video Fusion* branch in VF-Bench are shown in Fig. 12.
- Dataset visualizations for *Medical Video Fusion* branch in VF-Bench are shown in Fig. 13.

Our VF-Bench provides high-quality data in diverse scenes, serving as a strong benchmark for future video fusion tasks.

C Quantitative Results on Low-Resolution MEF and MFF Branches

Considering that inference on full-resolution videos may not be suitable for small devices or scenarios with limited computational resources, we additionally conduct experiments on low-resolution versions of the MEF (540p) and MFF (480p) datasets in Tab. 5. While the primary training and evaluation of both branches are performed on the 2K-resolution datasets in the main paper, the low-resolution results presented here serve as complementary evidence to assess the consistency of performance across different input scales. The supplementary results further demonstrate that our model can consistently produce high-quality fused videos.

D Additional Qualitative Fusion Comparison Results

We present more fusion visualization results in the figures below and on the project homepage videos:

- More qualitative comparisons for *Multi-exposure Video Fusion* results are shown in Fig. 14.
- More qualitative comparisons for *Multi-focus Video Fusion* results are shown in Fig. 15.
- More qualitative comparisons for *Infrared-Visible Video Fusion* results are shown in Fig. 16.
- More qualitative comparisons for *Medical Video Fusion* results are shown in Fig. 17.

The visual comparisons support our observation and conclusions: our UniVF method effectively preserves fine details and texture from the source images, while comprehensively integrating information from different settings or modalities to produce richly informative fused images. The videos further show that our results exhibit superior temporal consistency, with significantly less flickering and motion incoherence.

E Limitations

We assume well alignment between modalities in the input videos, and we have accordingly filtered the videos in VF-Bench. However, there are, although rarely, still a few frames (1 out of 300 frames) that are not aligned. With this assumption and trained on our well-filtered data, the model may

produce artifacts or ghosting in the fused video when encountering misaligned frames, as shown in Fig. 18. In future work, we plan to incorporate alignment modules into our UniVF to improve the robustness of the fusion process against misaligned input frames while maintaining the model’s efficiency and fusion quality.

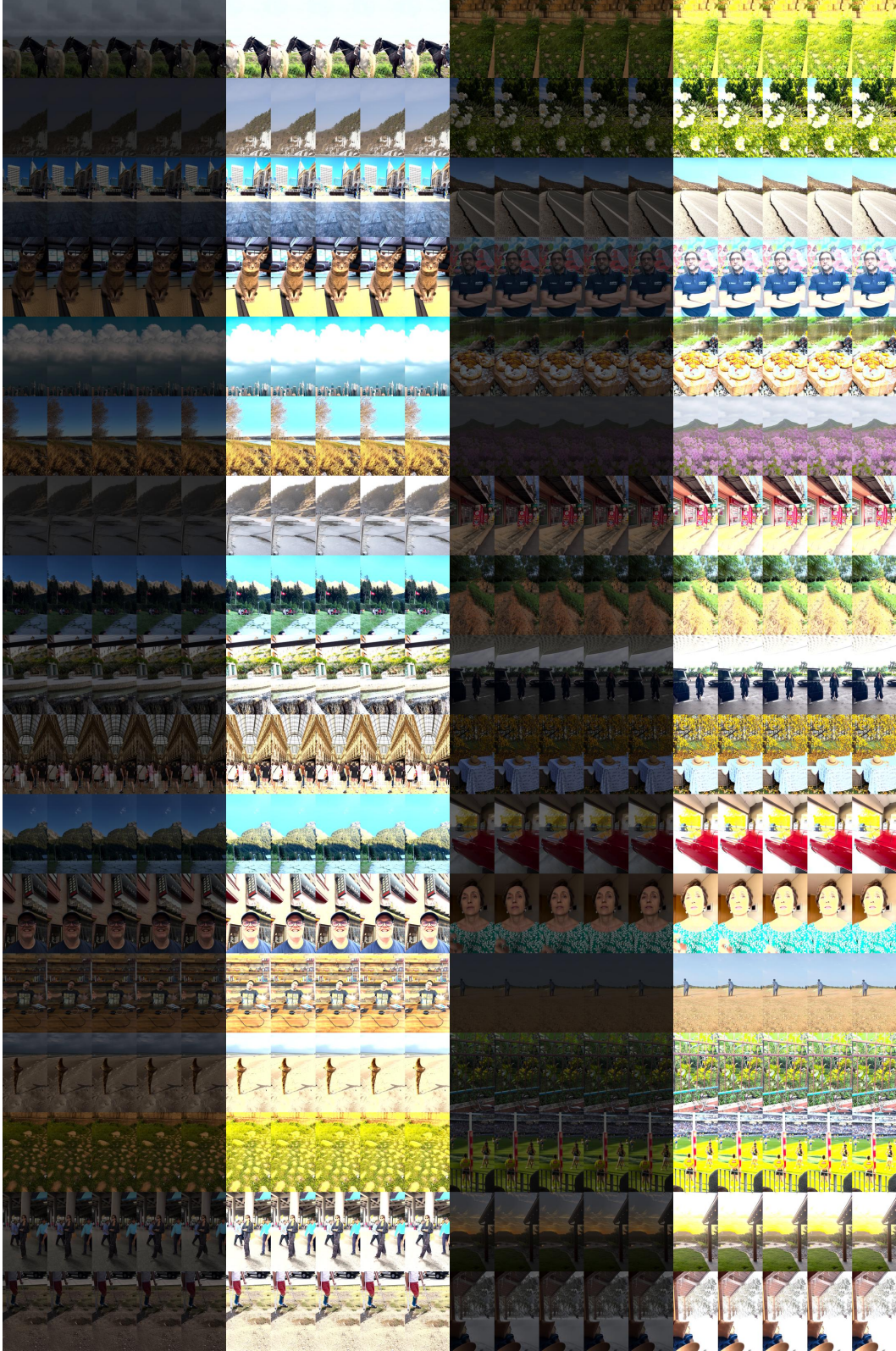


Figure 10: Dataset visualizations for *Multi-exposure Video Fusion* branch in VF-Bench. Columns 1–5 and 11–15 correspond to under-exposed video sequences, while columns 6–10 and 16–20 correspond to their respective over-exposed video sequences.



Figure 11: Dataset visualizations for *Multi-focus Video Fusion* branch in VF-Bench. Odd rows correspond to the far-focus video sequences and even rows correspond to the respective near-focus video sequences.

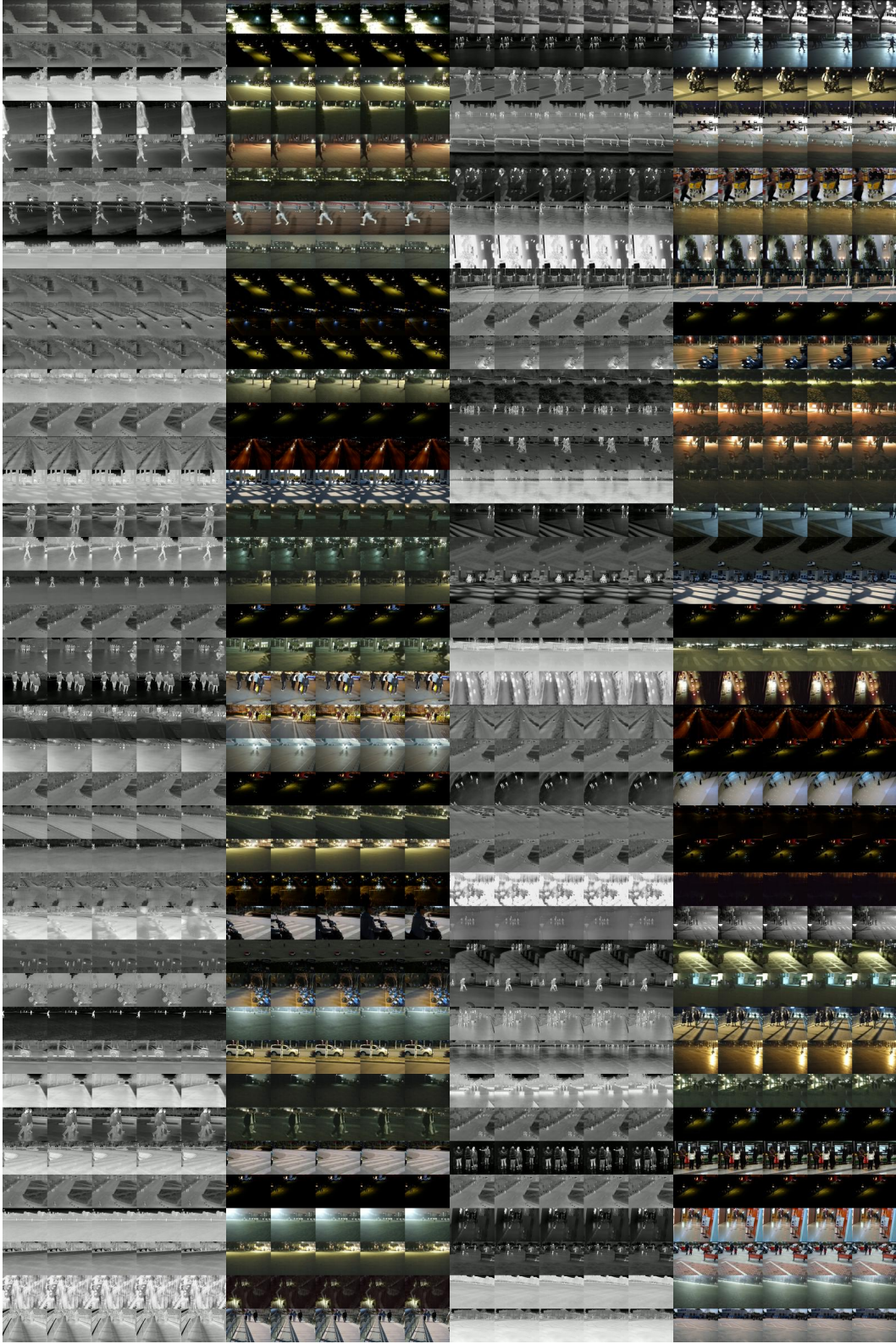


Figure 12: Dataset visualizations for *Infrared-Visible Video Fusion* branch in VF-Bench. Columns 1–5 and 11–15 correspond to infrared video sequences, while columns 6–10 and 16–20 correspond to their respective visible video sequences.

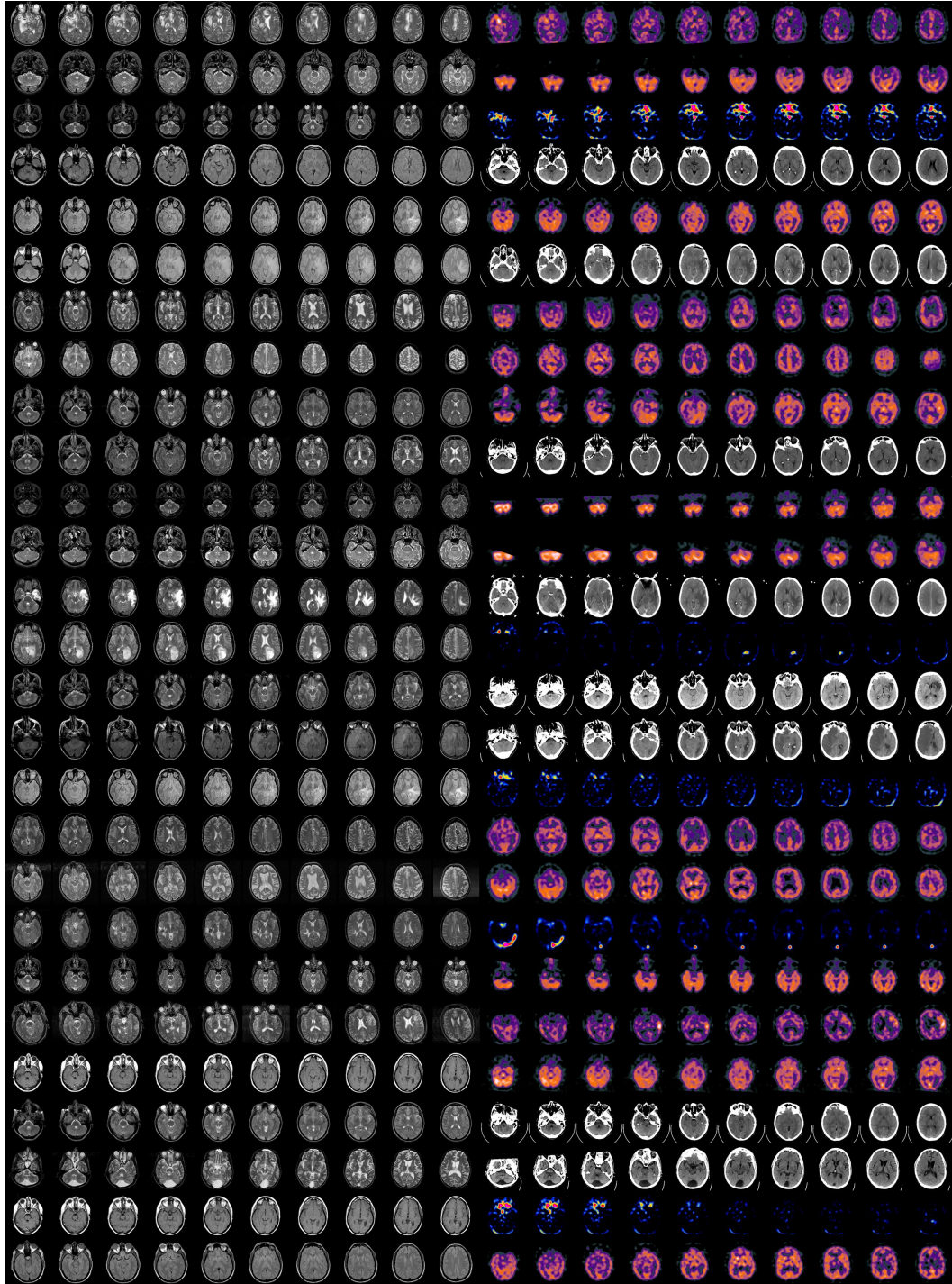


Figure 13: Dataset visualizations for *Medical Video Fusion* branch in VF-Bench. Columns 1–10 correspond to MRI video sequences, while columns 11–20 correspond to their respective CT, PET or SPECT video sequences.

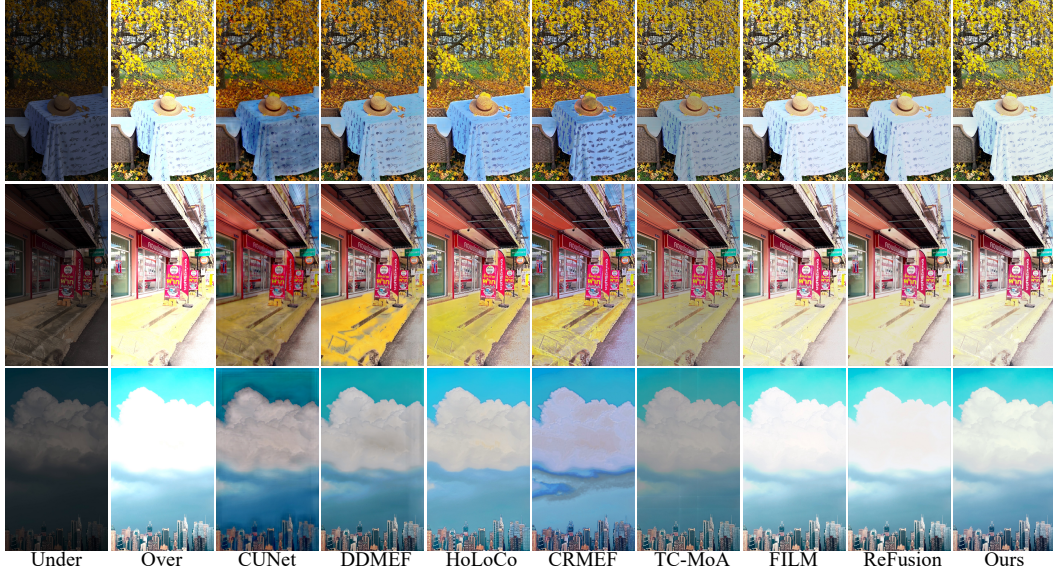


Figure 14: Visualization comparison of the fusion results in the multi-exposure video fusion task.

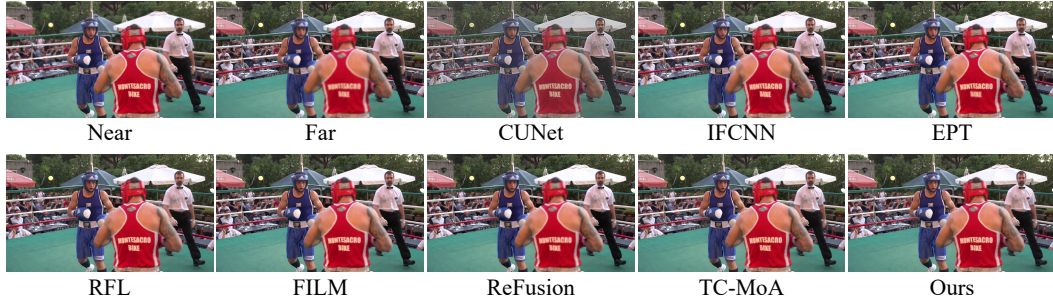


Figure 15: Visualization comparison of the fusion results in the multi-focus video fusion task.

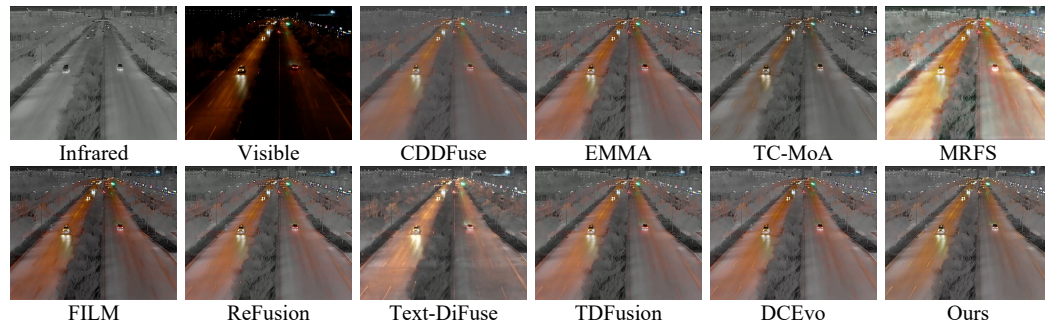


Figure 16: Visualization comparison of the fusion results in the infrared-visible video fusion task.

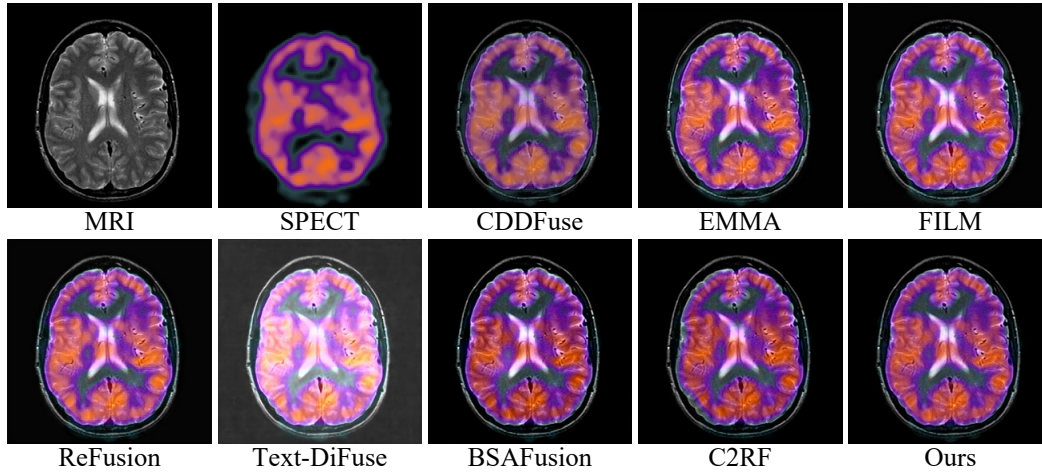


Figure 17: Visualization comparison of the fusion results in the medical video fusion task.



Figure 18: Artifacts in fused video caused by misaligned input frames.