Supplementary Material of "VideoFusion: A Spatio-Temporal Collaborative Network for Multi-modal Video Fusion"

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

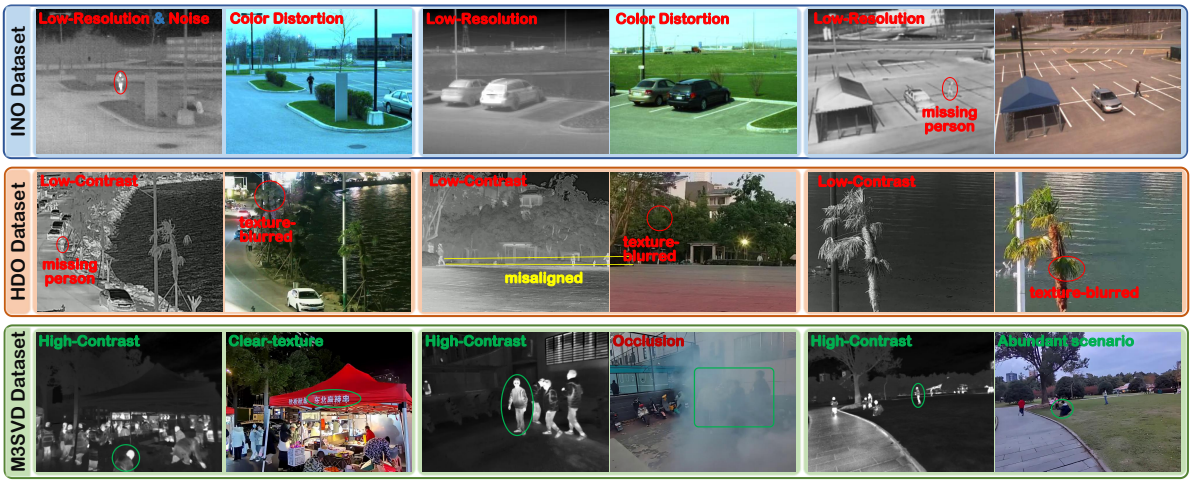

Figure s1. Comparative visualization of representative scenes from the HDO, INO, and our M3SVD datasets.

1. Comparative Analysis of Video Datasets

Table 1 compares the data scale and scene diversity across various multi-modal datasets. As noted, the TNO dataset is limited by its small data volume, the INO dataset suffers from low resolution, and the HDO dataset exhibits poor imaging quality, with all datasets covering only a narrow range of scenarios.

To substantiate these observations, we present source video frames from various video datasets in Fig. s1. As shown, the INO dataset is hampered by its low resolution, making it challenging to provide high-quality scene descriptions. Additionally, its infrared frames exhibit relatively low contrast, often failing to deliver information on significant objects. Moreover, its visible frames are prone to color distortion, with the first scene displaying a noticeable blue tint across the entire frame and the second scene showing a greenish hue. In the HDO dataset, poor imaging quality manifests as low-contrast infrared frames and blurred visible frames. Specifically, in the first scene, pedestrians in the infrared frames are barely discernible, and foliage in vis-

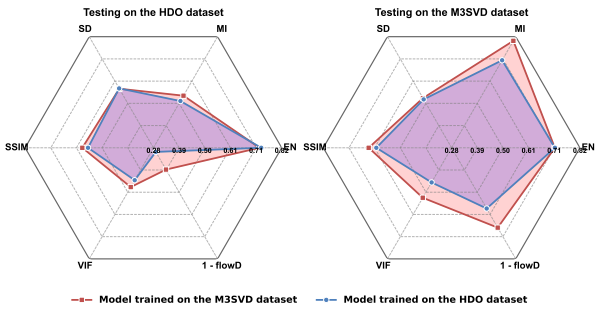

Figure s2. Performance of models trained on M3SVD and HDO when evaluated cross-dataset on HDO and M3SVD, respectively. All metrics are normalized to [0, 1] to eliminate scale differences.

*Corresponding author.

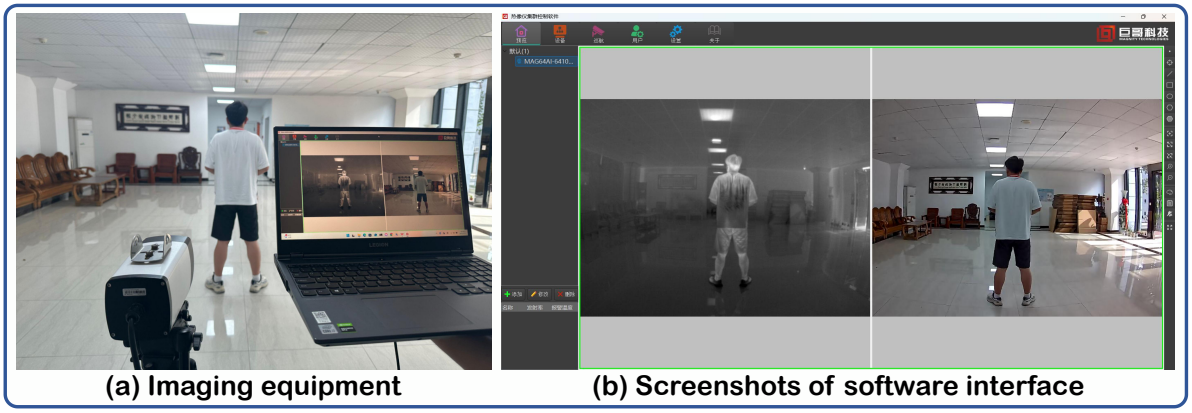

Figure s3. Schematic of our practical imaging device.

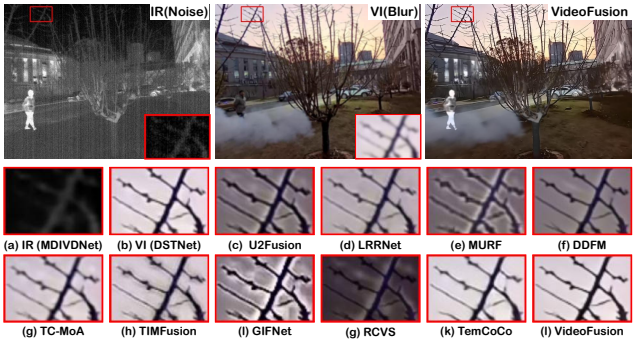

Figure s4. Fusion results under a challenging scenario.

ible frames appears blurred. In the second scene, a clear misalignment between the infrared and visible frames is evident. Furthermore, it is noteworthy that the HDO dataset contains numerous scenes lacking fusion value, where the infrared frames fail to provide meaningful information gain. In contrast, our proposed M3SVD dataset features high-contrast infrared video frames, clear visible video frames, and covers a diverse range of scenarios, including occlusion and camouflage, thereby offering robust support for validating the practical value of video fusion.

To further demonstrate the advantages of our M3SVD dataset, we train two models separately on the M3SVD and HDO datasets. The quantitative comparison results on the testing set of HDO and M3SVD datasets are presented in Fig. s2. Notably, when the model trained on M3SVD is directly generalized on the HDO dataset, it outperforms the model trained on the HDO dataset on several metrics, such as MI, SSIM, VIF, and flowD. Moreover, when both models are tested on the M3SVD dataset, the performance gap further widens. These observations indicate that the HDO dataset lacks sufficient scale, quality, and scene diversity to support strong cross-dataset generalization.

2. Dual-spectral Imaging System

Our M3SVD dataset is collected using a dual-spectrum sensor provided by Magnity Technology[1]. As shown in Fig. s3, the sensor is mounted on a tripod to prevent motion blur and spatial misalignment caused by camera shake. We employ proprietary acquisition software developed by Magnity Technology to enable real-time capture and synchronization of infrared and visible video streams. Fig. s3 (b) illustrates the raw data interface and examples of the recorded videos.

3. Supplementary Fusion Results Analysis

Figure s4 presents the fusion results under challenging scenarios. The source infrared video is affected by stripe noise, while the visible video suffers from blurring. Although MDIVDNet is capable of removing the stripe noise, it causes blurry representations of the tree branches. DSTNet enhances the sharpness of the branches but introduces artifacts to a certain extent. Such artifacts are particularly noticeable in U2Fusion, MURF, GIFNet, and RCVS. These methods amplify artifacts present in the visible frames and are also influenced by the degraded information from the blurred infrared frames. Although the video-based TemCoCo does not produce obvious artifacts, its ability to exploit complementary and temporal information to counter degradation is limited. As a result, its fusion results still suffer from noticeable blurring to some extent. In comparison, VideoFusion effectively exploits cross-modal complementary characteristics and temporal cues across dimensions to achieve better scene representations. It is capable of providing both sharp textures and enhanced prominent targets.

4. Analysis of Temporal Cue Effectiveness

As discussed previously, temporal cues are essential for maintaining consistency, a fundamental objective in video-

[1]https://www.magnity.com.cn/

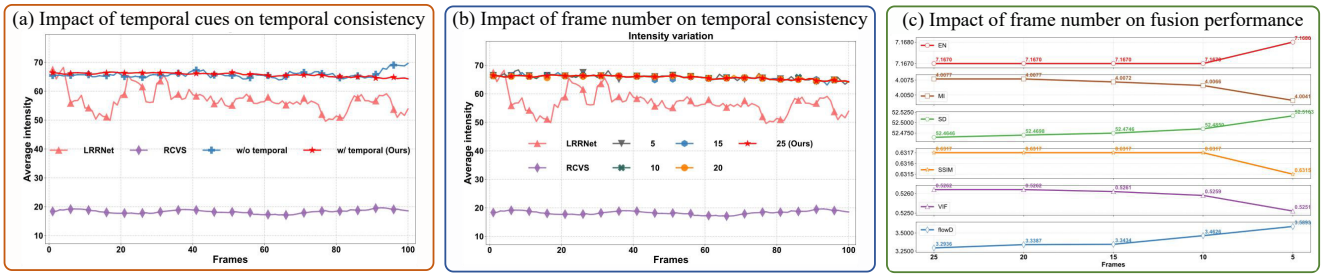

Figure s5. Impact of temporal cues and frame numbers.

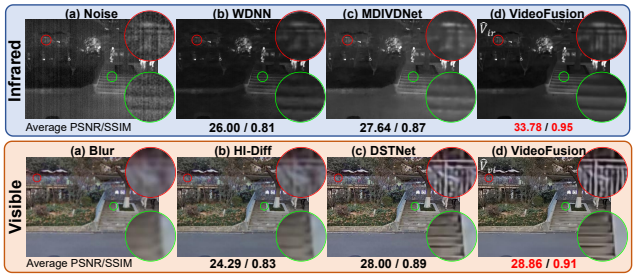

Figure s6. Visual comparison of ablation studies.

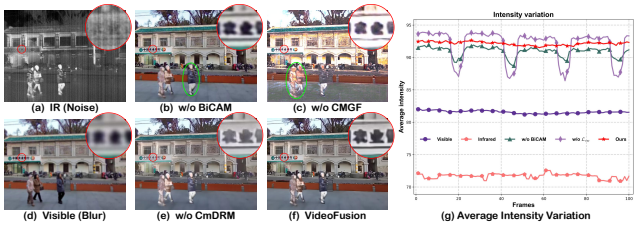

Figure s7. Qualitative and quantitative results on restoration.

related tasks. We demonstrate the effectiveness of temporal information in Fig. 6, 7, and 9. To further highlight the importance of temporal priors, we construct sequences by duplicating identical frames, thereby eliminating temporal cues. As shown in Fig. s5 (a), removing temporal cues in this manner leads to noticeable fluctuations in average frame brightness, indicating degraded temporal consistency and underscoring the critical role of temporal information.

Furthermore, we investigate the impact of varying frame counts on temporal consistency and fusion performance, as shown in Figs. s5 (b) and (c). Reducing the number of frames in the input sequence leads to noticeable inter-frame brightness fluctuations. However, when the frame count exceeds 10, these fluctuations become nearly imperceptible. Notably, we set the default frame count to 25 to maximize GPU utilization. Additionally, reducing the frame count marginally improves certain no-reference metrics, such as EN and SD, likely due to relaxed temporal consistency constraints. *Nevertheless, as the frame count decreases, the flowD metric, which reflects temporal consistency, shows a noticeable degradation, indicating that the temporal stability of the fused videos is compromised. This observation further underscores the importance of temporal modeling.* In addition, several reference-based metrics, including MI, SSSIM, and VIF, also exhibit slight performance drops.

5. Analysis of Modality Unmixing

As introduced earlier, our method incorporates a Modality Unmixing Unit designed to disentangle fused features into modality-specific components, enabling the reconstruction of corresponding infrared (IR) and visible (VI) video streams. While this module functions as an auxiliary branch, it effectively harnesses cross-modal complementary features and temporal cues to significantly enhance the quality of degraded inputs.

To validate its efficacy, we compare our restoration results against several baseline methods, including image-based approaches such as WDNN [3] and HI-Diff [2], as well as video-based methods, namely MDIVDNet [1] and DSTNet [4]. As illustrated in Fig. s6, our method achieves superior restoration performance, as evidenced by higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) scores averaged across the entire test set. Qualitatively, our approach excels in suppressing noise in IR video streams while successfully recovering critical structural details, such as staircases and fences. Moreover, the restored VI video streams produced by our VideoFusion framework exhibit sharper, more well-defined textures. These results demonstrate that the auxiliary restoration task not only guides the fusion network during training but also deepens its ability to comprehend scene structures under degraded conditions. Notably, video-based methods consistently outperform their image-based counterparts in both quantitative metrics and qualitative visual quality, underscoring the pivotal role of temporal information in advancing restoration performance.

6. Supplementary Ablation Study

Although the results in Fig. 9 already demonstrate the effectiveness of each proposed component, the quantitative metrics in Tab. 5 indicate that VideoFusion, equipped with all innovative components, exhibits shortcomings in the Spatial Frequency (SF) and Standard Deviation (SD) metrics. It is important to emphasize that SF and SD, as no-reference metrics, statistically evaluate the richness of texture details and contrast in the fusion results, respectively. When the fused video contains noticeable artifacts, as shown in Fig. 9 (b) and Fig. s7 (c), these metrics tend to be inflated. Notably, due to the absence of ground truth in multi-sensor fusion tasks,a comprehensive evaluation of algorithm effectiveness requires integrating qualitative and quantitative results, with quantitative analysis necessitating multiple metrics for a holistic assessment. To further highlight the importance of our key modules, we provide additional qualitative results in Fig. s7. It can be observed that our method delivers clearer text and more prominent target objects without being affected by artifacts. Additionally, our VideoFusion exhibits superior temporal consistency.

References

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

[2] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[3] Juntao Guan, Rui Lai, and Ai Xiong. Wavelet deep neural network for stripe noise removal. *IEEE Access*, 7:44544–44554, 2019. 3

[4] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22191–22200, 2023. 3