THERMAL SEQUENCE RECONSTRUCTION UNDER COUPLED MOTION BLUR AND ROLLING-SHUTTER READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic uncooled infrared imaging is affected by the coupled effects of thermal persistence and rolling-shutter (RS) readout, making fast scene dynamics difficult to recover. The non-instantaneous response of microbolometer pixels introduces an exponentially decaying temporal history into each observation, while row-dependent readout couples the resulting temporal blur with geometric distortion. Existing sequence reconstruction methods are designed for visible-light imaging, whereas infrared restoration methods generally target one-to-one recovery and do not jointly model these effects. In this work, we formulate thermal sequence reconstruction through a unified spatiotemporal observation model that captures the coupling between microbolometer thermal persistence and RS readout. Based on this formulation, we develop a physics-guided framework that reconstructs sharp scene states at target times from consecutive degraded observations and maps them to the target global-shutter (GS) geometry. During training, the reconstructed RS states are passed through the forward thermal response model to enforce physical reblur consistency. We further construct a physics-based synthetic benchmark and collect a real-world dataset using two uncooled infrared cameras. Experiments demonstrate improved performance over existing motion-restoration methods and favorable cross-sensor generalization on real uncooled infrared observations.

1 INTRODUCTION

Uncooled infrared cameras are widely used in robotics (Garcia et al., 2007; Huang et al., 2023), surveillance (Hasanujjaman et al., 2023; Niu et al., 2023a), and autonomous systems (Tang et al., 2022; Chen et al., 2024) owing to their low cost and power consumption. By sensing naturally emitted thermal radiation, they provide reliable observations under low illumination (Guo et al., 2020; Jiang et al., 2021; Niu et al., 2026a) and adverse weather, where visible-light imaging often becomes unreliable. However, images captured under motion often suffer from severe degradation, making fast scene dynamics difficult to perceive and analyze. Recovering the underlying scene evolution is therefore important for dynamic thermal imaging.

Beyond conventional deblurring, sequence reconstruction uses the temporal information in blurred observations to recover sharp frames over the observation interval, providing a time-resolved view of the underlying motion. For uncooled infrared cameras, however, the degradation encountered during dynamic acquisition extends beyond conventional motion blur. The non-instantaneous thermal response of microbolometer pixels introduces an exponentially decaying temporal history into each observation, differing from the finite and approximately uniform exposure integration commonly assumed in visible-light imaging. Meanwhile, rolling-shutter readout assigns different sampling times to different image rows, coupling temporal blur with geometric distortion under motion.

Existing sequence reconstruction methods (Oh & Kim, 2022; Zhong et al., 2023; Ji et al., 2023) are designed for visible-light imaging, which differs substantially from microbolometer imaging. Infrared motion-restoration methods, in contrast, generally recover a single sharp frame from each degraded observation. Among these methods, most adopt conventional blur formulations (Liu et al., 2026; Li et al., 2026), while a few explicitly model thermal blur caused by microbolometer persistence (Oswald-Tranta et al., 2010). Even the latter describe only the thermal response, leaving its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

interaction with row-dependent rolling-shutter sampling unmodeled. TRNeRF (Carmichael et al., 2025) instead models these coupled degradations through scene-level radiance-field reconstruction and multi-view constraints. However, direct temporally resolved restoration from degraded infrared videos without relying on a scene-level 3D representation remains largely unexplored.

To address this gap, we formulate thermal sequence reconstruction using a unified observation model that couples thermal persistence with rolling-shutter (RS) readout. Building on this model, we develop a physics-guided framework that reconstructs sharp scene states at target times from consecutive degraded observations. Specifically, using the center RS frame as the reference grid, we align and fuse the input observations to aggregate cross-frame information. From the fused representation, a time-conditioned state-space reconstruction module with wavelet-based refinement recovers sharp states at the target times on this reference grid. Motion-guided geometric rectification then maps each reconstructed state to the corresponding global-shutter (GS) geometry. In parallel, an auxiliary RS decoder reconstructs sharp RS states at the same target times, which are passed through the forward thermal response model to enforce physical reblur consistency during training. For controlled evaluation and real-world validation, we construct a physics-based synthetic benchmark from high-frame-rate sharp infrared sequences captured with a cooled infrared camera and collect a real-world dataset using two uncooled infrared cameras. The main contributions are summarized as follows:

- We formulate thermal sequence reconstruction for uncooled infrared imaging using a unified spatiotemporal observation model that couples microbolometer thermal persistence with rolling-shutter readout.
- Based on this formulation, we develop a physics-guided restoration framework that reconstructs sharp scene states at target times and further rectifies their readout-dependent geometry for sharp global-shutter recovery.
- We introduce a forward imaging consistency constraint that jointly constrains all predicted RS states through the thermal response model, explicitly linking temporally resolved restoration to the physical formation of the observed blur.
- We introduce a physically synthesized benchmark and a real-world dataset from two uncooled infrared cameras. Experiments show improved performance over existing motion restoration pipelines and favorable cross-sensor generalization on real data.

2 Related Work

Motion from Blur. Traditional motion deblurring restores a sharp image from each blurred observation. Related multi-frame restoration methods exploit neighboring frames through spatially adaptive filtering (Zhou et al., 2019), deformable alignment (Jiang et al., 2022), recurrent temporal modeling (Zhong et al., 2020), flow-guided propagation (Lin et al., 2022; Zhang et al., 2024), deformable temporal fusion (Niu et al., 2023c), or motion-aware Gaussian rendering (Niu et al., 2026b). Beyond such one-to-one restoration, sequence reconstruction uses the temporal information in blurred observations to recover sharp frames over the observation interval. Existing approaches combine deblurring and interpolation through cascaded (Jin et al., 2019), recurrent (Shen et al., 2020), or flow-guided (Oh & Kim, 2022) reconstruction. More recent methods support arbitrary-time recovery (Zhong et al., 2023) or jointly address motion blur and rolling-shutter distortion in video reconstruction (Ji et al., 2023).

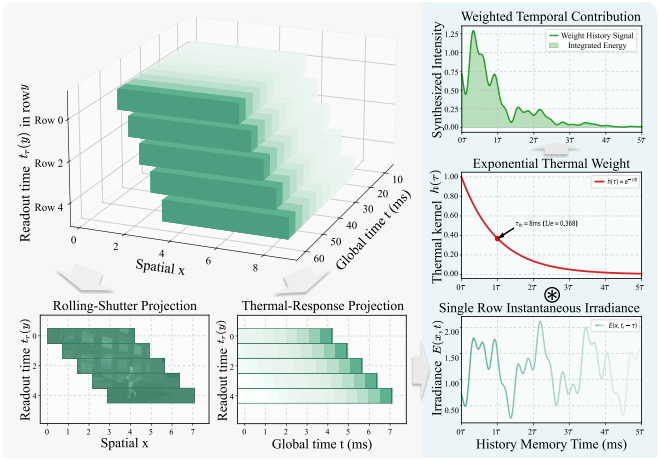
Motion from Thermal Blur in Uncooled Infrared Imaging. Existing restoration methods for motion-degraded infrared imagery have largely focused on one-to-one deblurring rather than temporally resolved sequence reconstruction. These methods address motion blur through spatial kernel estimation (Varghese et al., 2022), learned representations (Yi et al., 2023), or cross-frame motion cues (Yu et al., 2026). Physics-based studies more explicitly model thermal persistence using exponential response models (Oswald-Tranta et al., 2010), temporal inversion (Ramanagopal et al., 2020), inertial guidance (Lee et al., 2022), or microbolometer-aware synthesis (Ko et al., 2023), but still focus on thermal-blur restoration rather than sequence reconstruction under the coupled degradations. TRNeRF (Carmichael et al., 2025) incorporates thermal persistence, rolling-shutter readout, and sensor noise into scene-level 3D reconstruction, but requires scene-specific multi-view optimization. Direct feed-forward sequence reconstruction from degraded infrared videos therefore remains largely unexplored.

Rolling Shutter Correction. Rolling shutter correction recovers undistorted global-shutter images from RS observations. Existing methods are broadly divided into single-frame and multi-frame approaches. Single-frame methods rely on geometric or motion priors, including joint depth-motion estimation (Zhuang et al., 2019) and homography-based modeling (Yan et al., 2023), whereas multi-frame methods exploit inter-frame correspondences through multi-hypothesis alignment (Cao et al., 2022), nonlinear trajectory modeling (Qu et al., 2023), or bidirectional correction with diffusion-based occlusion recovery (Yang et al., 2025). RS-NeRF (Niu et al., 2024) jointly recovers camera trajectory and GS radiance fields from RS videos.

3 METHOD

3.1 PHYSICS-GUIDED SPATIOTEMPORAL FORMULATION

Dynamic uncooled infrared imaging is jointly affected by motion blur induced by thermal persistence and geometric distortion from rolling-shutter readout. Rather than modeling these degradations independently, we formulate their interaction within a unified image formation process. Thermal persistence and row-wise readout are coupled through time: each sensor row is sampled at a distinct instant while retaining a decaying response to earlier irradiance. Fig. 1 (a) visualizes this interaction as a spatiotemporal observation volume, whose orthogonal projections expose the corresponding rolling-shutter displacement and thermal persistence. Specifically, let $L(t)$ denote the latent instantaneous irradiance image at time t, with $L(x, y, t)$ representing its value at pixel (x, y). Unlike conventional visible-light imaging (Niu et al., 2023b), where motion blur is commonly modeled as integration over a finite exposure interval, a microbolometer continuously retains an exponentially decaying response to previously received radiation (Carmichael et al., 2025). Assuming all pixels are synchronously read out at time t, the corresponding noise-free thermal response is modeled as:



(a) Spatiotemporal Imaging Model (b) Thermal Blur Formation

Figure 1: Unified spatiotemporal imaging process of an uncooled infrared rolling-shutter camera.

$$I_{\text{th}}(t) = \frac{1}{\tau_{\text{th}}} \int_{-\infty}^{t} \exp\left(\frac{s-t}{\tau_{\text{th}}}\right) L(s)\, ds, \tag{1}$$

where τ_{th} is the thermal time constant. Thus, all pixels in $I_{\text{th}}(t)$ share the same temporal endpoint t. In practice, we truncate the thermal history to $T_m = 5\tau_{\text{th}}$, which retains over 99% of the response. Fig. 1 (b) illustrates the temporal accumulation for a single row. In addition, rolling-shutter readout breaks this global synchronization by assigning different readout times to different rows. For an image of height H, the readout time of row y is:

$$t_r(y) = t_c + \left(\frac{y}{H-1} - \frac{1}{2}\right) T_r, \tag{2}$$

where T_r is the frame readout duration and t_c denotes the readout time of the center row. Consequently, each row has a different temporal endpoint for its thermal history. Combining the row-dependent readout with the thermal response yields the observed frame:

$$I(x, y) = \frac{1}{\tau_{\text{th}}} \int_{t_r(y)-T_m}^{t_r(y)} \exp\left(\frac{s - t_r(y)}{\tau_{\text{th}}}\right) L(x, y, s)\, ds + n(x, y), \tag{3}$$

where $n(x, y)$ denotes additive sensor noise, including structured banding arising from pixel and readout-channel nonuniformity. Based on the coupled imaging process in Eq. 3, we formulate

restoration as a factorized process that separates temporal-state recovery from readout-induced geometric correction, while preserving consistency with the forward imaging model during training. As shown in Fig. 2, given three consecutive degraded observations, motion-aware temporal alignment registers the neighboring-frame features to the center-frame reference, after which feature fusion aggregates the aligned features. The fused feature is then transformed into a time-conditioned latent representation for each queried time on the center RS grid. A geometry-guided branch further maps this representation to the target GS frame by estimating row-dependent correspondences. In parallel, an auxiliary RS decoder reconstructs RS images at the queried times to provide the RS-domain states required by the forward imaging model. These states are then passed through the thermal response model to reproduce the degraded center observation. Motion-aware temporal alignment and time-conditioned state reconstruction are detailed in Sec. 3.2, followed by geometric rectification in Sec. 3.3 and forward imaging consistency in Sec. 3.4.

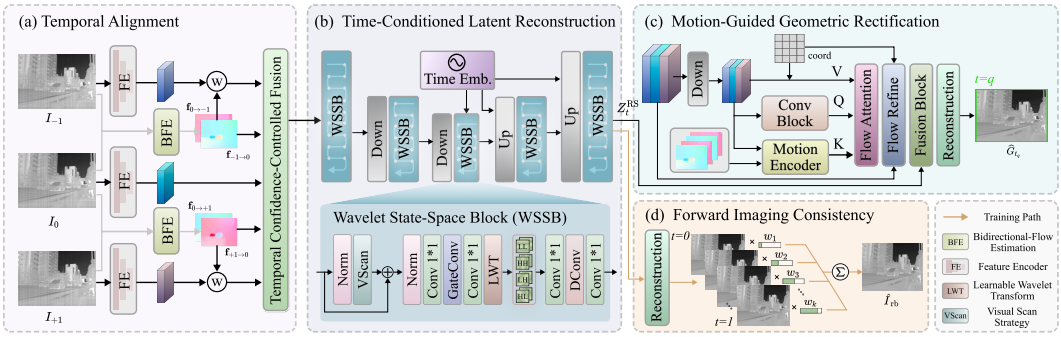


Figure 2: Overview of the proposed physics-guided framework for thermal sequence reconstruction.

3.2 Time-Conditioned Latent Reconstruction

Given three consecutive degraded observations $\{I_{-1}, I_0, I_{+1}\}$, we first construct a temporally queryable latent representation on the center RS grid. A shared feature encoder (FE) extracts the frame-wise features F_{-1}, F_0, and F_{+1}, while a bidirectional flow estimator (BFE) (Teed & Deng, 2020) estimates two pairs of bidirectional motion fields between the center frame and its neighboring frames. Thermal blur tends to weaken local structures in individual infrared observations, whereas neighboring frames provide complementary temporal and structural information. For center-referenced temporal aggregation, we use the neighbor-to-center flows $\mathbf{f}_{-1\to0}$ and $\mathbf{f}_{+1\to0}$ to warp the neighboring features onto the center RS grid, while the reverse flows are retained for subsequent geometric rectification. Since optical flow estimated from blurred infrared observations can be unreliable, the proposed Temporal Confidence-Controlled Fusion Module (TCCF) adopts a center-dominant strategy to suppress contributions from erroneously aligned neighboring features:

$$F_{\text{agg}} = \mathcal{A}\left(F_0, \mathcal{W}(F_{-1}, \mathbf{f}_{-1\to0}), \mathcal{W}(F_{+1}, \mathbf{f}_{+1\to0})\right), \tag{4}$$

where $\mathcal{W}(\cdot)$ denotes feature warping, and $\mathbf{f}_{-1\to0}$ and $\mathbf{f}_{+1\to0}$ are the motion fields estimated by the BFE. $\mathcal{A}(\cdot)$ preserves the center feature as an identity path and injects each aligned neighboring feature only through a gated residual. We next derive a time-conditioned representation for each queried state on the center RS grid using a state-space reconstruction module composed of multiple Wavelet State-Space Blocks (WSSBs). As illustrated in Fig. 2, each WSSB first performs LayerNorm followed by a Visual Scan (VScan) and residual aggregation. Following Kong et al. (2025), VScan employs a selective state-space scan with a fixed traversal direction. To propagate information along complementary spatial orientations, we alternate flip and transpose operations across successive WSSBs, mapping different orientations onto the same scanning operator without introducing multiple scans within each block. Such long-range and orientation-aware propagation facilitates the recovery of spatially extended structures degraded by thermal persistence.

The scanned feature is subsequently normalized and processed by 1×1 projections and GateConv, which multiplicatively gates the normalized feature using its depthwise-convolved response. We further introduce wavelet-domain refinement to separate frequency and directional components. The feature is decomposed through a learnable wavelet transform initialized with Haar filters into $\{\text{LL}, \text{LH}, \text{HL}, \text{HH}\}$ subbands, followed by channel mixing and depthwise convolution before inverse wavelet reconstruction. Since thermal persistence suppresses transient spatial details and

produces direction-dependent smearing under motion, the wavelet subbands separately represent slowly varying thermal structures and directional high-frequency details. To reconstruct different latent states from the same degraded observations, the early feature extraction remains independent of the queried time, while the temporal query is represented by a normalized scalar $t \in [0, 1]$ and injected into the subsequent reconstruction stages through TimeFiLM:

$$\mathrm{FiLM}(X, t) = X \odot (1 + \gamma(t)) + \beta(t), \tag{5}$$

where $\gamma(t)$ and $\beta(t)$ are generated from the temporal query. The shared representation is therefore reused across different temporal queries, while varying t produces the corresponding time-conditioned latent representation Z_t^{RS} within the thermal history. The resulting Z_t^{RS} encodes the sharp scene state at the queried time while remaining on the center RS grid, and is subsequently passed to geometric rectification.

3.3 Motion-Guided Geometric Rectification

The representation Z_t^{RS} recovered in Sec. 3.2 contains sharp scene content at the queried time t, but still follows the row-dependent RS geometry. We therefore estimate GS-to-RS distortion fields and geometrically map Z_t^{RS} onto the target GS grid. As illustrated in Fig. 2, the frame-wise features extracted by the shared FE are processed by a lightweight convolutional pyramid to construct multi-scale geometric representations. To preserve the native RS geometry, the rectification branch uses the frame-wise features before temporal alignment rather than the center-aligned feature F_{agg}. For a temporal query $t \in [0, 1]$, the row readout times are expressed in the same normalized temporal coordinate as t, yielding the normalized row-time coordinate $\bar{t}_r(y)$. We then define the row-time offset map as $T_t(y) = \bar{t}_r(y) - t$ and use it together with the bidirectional motion fields from the BFE to condition correspondence estimation.

Following the distortion-flow formulation (Cao et al., 2024), we construct a motion-guided Flow Attention module to establish coarse GS-to-RS correspondences. At the coarsest scale, the three geometric features are fused with the motion and temporal conditions to form the query Q. For each input frame, its geometric feature is projected to an appearance key augmented by encoded bidirectional motion, while the value concatenates the geometric feature with normalized RS-grid coordinates. The attention therefore retrieves a matched feature $\widetilde{F}_i$ and its corresponding RS-grid location $\widetilde{P}_i$ for each target GS position. Their coordinate difference $\Delta P_i = \widetilde{P}_i - P$ is combined with $\widetilde{F}_i$ to form the correspondence representation. The three correspondence representations are then fused with the motion representation, from which a flow head predicts three initial GS-to-RS distortion fields.

The initial fields are progressively refined by the multi-scale FlowRefine module. At each level, the upsampled fields warp the corresponding geometric features, which are fused with higher-resolution geometry, motion, and temporal conditions. A flow head then directly re-estimates the complete distortion fields at the current scale. Repeating this procedure yields a full-resolution field for each input frame. Finally, the refined center-frame field $\mathbf{u}_t$ is applied to the time-conditioned representation: $\widetilde{Z}_t^{\mathrm{GS}} = \mathcal{W}\left(Z_t^{\mathrm{RS}}, \mathbf{u}_t\right)$, where $\mathcal{W}(\cdot)$ denotes backward warping. Large RS displacement may map part of the target GS grid outside the spatial support of the center observation. We therefore preserve the center-frame reconstruction in supported regions and introduce warped neighboring features through the Fusion block only for out-of-bounds regions. The resulting representation is finally decoded into the sharp global-shutter frame $\widehat{G}_t$.

3.4 Forward Imaging Consistency

To enforce consistency between the reconstructed states and the degraded observation under its RS sampling geometry, we incorporate the forward thermal image formation process into training. Let $\mathcal{H}$ denote the set of discrete sampling times spanning the truncated thermal history. The subset $\mathcal{Q} \subset \mathcal{H}$ contains the K most recent query times, whose associated states receive the dominant response weights. The reblurred observation is synthesized as $\widehat{I}_{\mathrm{rb}} = \sum_{t \in \mathcal{Q}} w_t \widehat{R}_t + \sum_{t \in \mathcal{H} \setminus \mathcal{Q}} w_t R_t$, where w_t is the discretized exponential-response weight at time t, derived from Eq. 1. For each $t \in \mathcal{Q}$, the shared reconstruction branch produces Z_t^{RS} conditioned on t, which the shared auxiliary RS decoder maps to the corresponding sharp RS state $\widehat{R}_t$. For each remaining time $t \in \mathcal{H} \setminus \mathcal{Q}$, R_t is taken from

the sharp reference sequence. Together, the reconstructed and reference states cover the complete truncated thermal history in the forward synthesis. The forward imaging consistency loss is then defined as $\mathcal{L}_{\text{phy}} = \left\| \widehat{I}_{\text{rb}} - \bar{I}_0 \right\|_1$, where $\bar{I}_0$ denotes the noise-free degraded center observation. While $\mathcal{L}_{\text{phy}}$ constrains the weighted aggregation of the reconstructed states, $\mathcal{L}_{\text{RS}}$ provides direct state-wise supervision: $\mathcal{L}_{\text{RS}} = \frac{1}{K} \sum_{t \in \mathcal{Q}} \left\| \widehat{R}_t - R_t \right\|_1$. For GS-domain supervision, we randomly sample one $t_q \in \mathcal{Q}$ at each iteration, rectify the corresponding representation, and supervise the resulting GS output: $\mathcal{L}_{\text{GS}} = \left\| \widehat{G}_{t_q} - G_{t_q} \right\|_1$, where G_{t_q} denotes the corresponding sharp GS reference. The overall reconstruction objective is defined as $\mathcal{L}_{\text{rec}} = \lambda_{\text{phy}} \mathcal{L}_{\text{phy}} + \lambda_{\text{RS}} \mathcal{L}_{\text{RS}} + \lambda_{\text{GS}} \mathcal{L}_{\text{GS}}$, where λ_{phy}, λ_{RS}, and λ_{GS} balance the three supervision terms.

Figure 3: Infrared cameras used for dataset construction. (a) captures high-frame-rate sharp sequences for the synthetic benchmark, while (b) and (c) are used to construct the real-world benchmark.



(a) X6981-HS SLS-PH (b) VIM384GS-USL (c) HM-TM06-LF/ANT

4 The Proposed Dataset

To evaluate our method under controlled and real imaging conditions, we construct a synthetic benchmark with supervision and a real-world dataset captured by two uncooled infrared cameras.

Synthetic Dataset. We build the synthetic benchmark from sharp infrared sequences captured by a cooled FLIR X6981-HS SLS-PH global-shutter camera (Fig. 3 (a)) at 1,000 fps and 640×512 resolution. To provide sufficient temporal resolution for rolling-shutter synthesis, the sequences are temporally interpolated by $8\times$ using FLAVR (Kalluri et al., 2023; Veit et al., 2026). Degraded observations are then generated using the unified imaging model in Sec. 3.1, with a rolling-shutter readout duration of $T_r = 32$ ms, a thermal time constant of $\tau_{\text{th}} = 8$ ms, representative of typical uncooled microbolometer responses, and a truncated thermal history of $T_m = 5\tau_{\text{th}}$ (Carmichael et al., 2025). The thermal integral is discretized every 2 ms into 21 exponential taps. For each observation, the GS reference time is aligned with the center-row readout, while the RS target follows the row-dependent readout without thermal integration. Sensor noise (Liu et al., 2026) is further added to the degraded observations. Finally, the synthetic dataset contains 9,015 training and 1,038 test observations.

Real-World Dataset. We further construct a real-world benchmark using two uncooled LWIR cameras from different manufacturers. The Vision Sensing VIM384GS-USL (Fig. 3(b)) employs a 384×288 detector with a 17 μm pixel pitch and an 8–14 μm spectral range, and has a specified thermal time constant below 10 ms. The HIKMICRO HM-TM06-LF/ANT (Fig. 3(c)) uses a 640×512 detector with the same 17 μm pixel pitch and 8–14 μm spectral range. We capture infrared videos under diverse camera and scene motions, naturally exhibiting thermal blur and rolling-shutter distortion. Since sharp GS ground truth is unavailable, these sequences are used for qualitative evaluation and cross-sensor generalization assessment. The benchmark contains 300 frames from the VIM384GS-USL and 200 frames from the HM-TM06-LF/ANT.

5 Experiments

Implementation Details. We implement our method in PyTorch and train it on three NVIDIA RTX 3090 GPUs. The time-conditioned RS reconstruction branch is first trained for 90 epochs, followed by 10 epochs of geometric rectification with the RS branch frozen and 40 epochs of joint fine-tuning. We use AdamW with an initial learning rate of 1×10^{-4}, weight decay 1×10^{-4}, and a batch size of 8 with 256×256 crops. Each degraded observation is synthesized from 21 temporal states. We set $K = 6$ and reconstruct the six most recent states, which together account for approximately 78% of the normalized truncated thermal-response weight under our discretization. The loss weights are set to $\lambda_{\text{RS}} = 1$, $\lambda_{\text{GS}} = 1$, and $\lambda_{\text{phy}} = 0.1$.

Experimental Setting. Our task jointly involves motion deblurring, rolling-shutter correction (RSC), and temporal interpolation. Since most existing methods address only part of this pipeline,

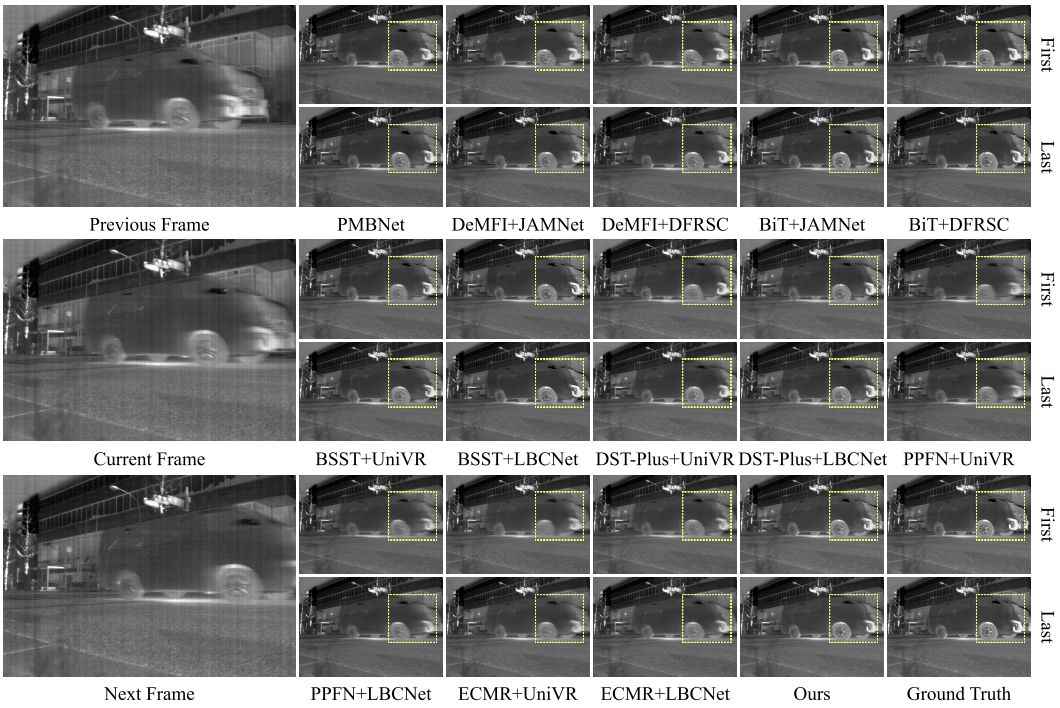


Figure 4: Visual comparison on the synthetic dataset. For each method, the top and bottom rows show the first and last reconstructed frames among the target times.

Table 1: Quantitative comparison on the proposed synthetic dataset. The best results are highlighted in red, with relative differences shown in blue.

Method	*All Query Times*			*Final Query*		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PMBNet	$26.1532_{\downarrow 9.9\%}$	$0.9221_{\downarrow 3.0\%}$	$0.1538_{\uparrow 35.6\%}$	$25.9677_{\downarrow 10.7\%}$	$0.9201_{\downarrow 1.4\%}$	$0.1492_{\uparrow 32.6\%}$
DeMFI+JAMNet	$27.6313_{\downarrow 4.0\%}$	$0.9238_{\downarrow 2.8\%}$	$0.1581_{\uparrow 37.3\%}$	$27.4652_{\downarrow 4.7\%}$	$0.9196_{\downarrow 1.4\%}$	$0.1532_{\uparrow 34.3\%}$
DeMFI+DFRSC	$28.2686_{\downarrow 1.7\%}$	$0.9319_{\downarrow 1.9\%}$	$0.1521_{\uparrow 34.8\%}$	$27.4137_{\downarrow 4.9\%}$	$0.9212_{\downarrow 1.2\%}$	$0.1494_{\uparrow 32.7\%}$
BiT+JAMNet	$26.6203_{\downarrow 8.0\%}$	$0.9305_{\downarrow 2.1\%}$	$0.1244_{\uparrow 20.3\%}$	$26.6113_{\downarrow 8.1\%}$	$0.9288_{\downarrow 0.4\%}$	$0.1207_{\uparrow 16.7\%}$
BiT+DFRSC	$27.7170_{\downarrow 3.7\%}$	0.9499	$0.1111_{\uparrow 10.8\%}$	$26.5945_{\downarrow 8.1\%}$	$0.9312_{\downarrow 0.2\%}$	$0.1131_{\uparrow 11.1\%}$
BSST+UniVR	$26.8418_{\downarrow 7.1\%}$	$0.9192_{\downarrow 3.3\%}$	$0.1215_{\uparrow 18.4\%}$	$26.5479_{\downarrow 8.3\%}$	$0.9144_{\downarrow 2.0\%}$	$0.1232_{\uparrow 18.3\%}$
BSST+LBCNet	$26.1521_{\downarrow 9.9\%}$	$0.9122_{\downarrow 4.1\%}$	$0.1302_{\uparrow 23.9\%}$	$26.1031_{\downarrow 10.2\%}$	$0.9140_{\downarrow 2.0\%}$	$0.1272_{\uparrow 20.9\%}$
DST-Plus+UniVR	$27.4227_{\downarrow 4.8\%}$	$0.9123_{\downarrow 4.1\%}$	$0.1553_{\uparrow 36.2\%}$	$27.5415_{\downarrow 4.4\%}$	$0.9110_{\downarrow 2.4\%}$	$0.1536_{\uparrow 34.5\%}$
DST-Plus+LBCNet	$27.2997_{\downarrow 5.3\%}$	$0.9140_{\downarrow 3.9\%}$	$0.1585_{\uparrow 37.5\%}$	$27.7164_{\downarrow 3.8\%}$	$0.9210_{\downarrow 1.3\%}$	$0.1508_{\uparrow 33.3\%}$
PPFN+UniVR	$26.1365_{\downarrow 10.0\%}$	$0.9169_{\downarrow 3.6\%}$	$0.1374_{\uparrow 27.9\%}$	$25.3272_{\downarrow 13.5\%}$	$0.9028_{\downarrow 3.3\%}$	$0.1431_{\uparrow 29.7\%}$
PPFN+LBCNet	$25.7401_{\downarrow 11.7\%}$	$0.9147_{\downarrow 3.8\%}$	$0.1406_{\uparrow 29.5\%}$	$25.1268_{\downarrow 14.4\%}$	$0.9085_{\downarrow 2.7\%}$	$0.1410_{\uparrow 28.7\%}$
ECMR+UniVR	$27.3682_{\downarrow 5.0\%}$	$0.8989_{\downarrow 5.7\%}$	$0.1902_{\uparrow 47.9\%}$	$27.1432_{\downarrow 5.9\%}$	$0.8945_{\downarrow 4.3\%}$	$0.1896_{\uparrow 46.9\%}$
ECMR+LBCNet	$27.0373_{\downarrow 6.3\%}$	$0.8990_{\downarrow 5.7\%}$	$0.1927_{\uparrow 48.6\%}$	$26.9353_{\downarrow 6.8\%}$	$0.8998_{\downarrow 3.6\%}$	$0.1868_{\uparrow 46.1\%}$
Ours	28.7428	$0.9334_{\downarrow 1.8\%}$	0.0991	28.7565	0.9326	0.1006

we construct complete baselines by combining complementary methods when necessary. We compare with the full-pipeline method PMBNet (Ji et al., 2023), and combine sequence reconstruction methods DeMFI (Oh & Kim, 2022) and BiT (Zhong et al., 2023) with multi-frame RSC methods JAMNet (Fan et al., 2023) and DFRSC (Cao et al., 2024). We further pair video deblurring methods BSST (Zhang et al., 2024) and DST-Plus (Pan et al., 2025) with rolling-shutter interpolation methods UniVR (Fan et al., 2024b) and LBCNet (Fan et al., 2024a). To further evaluate infrared-specific restoration, we also combine the infrared restoration methods PPFN (Liu et al., 2026) and ECMR (Li et al., 2026) with UniVR and LBCNet. In total, 13 competing methods/pipelines are evaluated. For fair comparison, all competing methods are retrained on our training set using their official implementations. We evaluate the reconstructed sharp states using PSNR, SSIM, and LPIPS (Zhang et al., 2018). For each degraded observation, the metrics are computed at K query times against their corresponding time-aligned ground truths and averaged over the query times, reported as *All Query Times*. We also report the most recent queried state as the *Final Query*. Owing to causal thermal persistence, this state contributes most strongly to the observation and thus directly reflects the deblurring capability.

5.1 Comparison with State-of-the-Art Methods

Fig. 4 presents qualitative comparisons on a challenging driving scene with fast vehicle motion, where motion blur and rolling-shutter distortion occur simultaneously. For each method, we visualize the reconstructed first and last sharp frames among the queried target times. In the yellow zoom-in region, the moving wheel and background line structures jointly reveal the ability to recover sharp details while correcting RS distortion. Among the competing methods, only BSST+LBCNet and DST-Plus+LBCNet recover noticeable wheel details across both displayed queried states, while most competing methods suffer from substantial detail loss. Although DST-Plus+LBCNet, BiT+JAMNet, and DeMFI+JAMNet partially rectify the wheel geometry, they introduce visible deformation in the originally straight background structures. In contrast, our method consistently recovers clearer wheel details, corrects the RS distortion, and better preserves the background geometry. The quantitative results in Tab. 1 further support the above observations. Our method achieves the best performance on all reported metrics except for the average SSIM over all queried times, where BiT+DFRSC obtains a marginally higher score. Overall, the quantitative advantage is consistent with the visual comparisons, demonstrating the effectiveness of our unified framework in jointly recovering sharp temporal details and correcting rolling-shutter distortion.

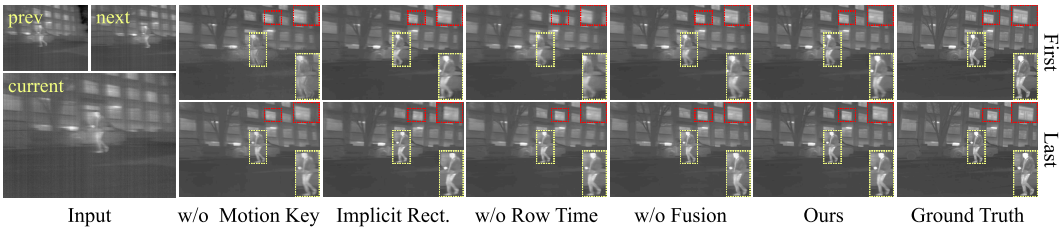


Figure 5: Visual comparison for the ablation of geometric rectification.

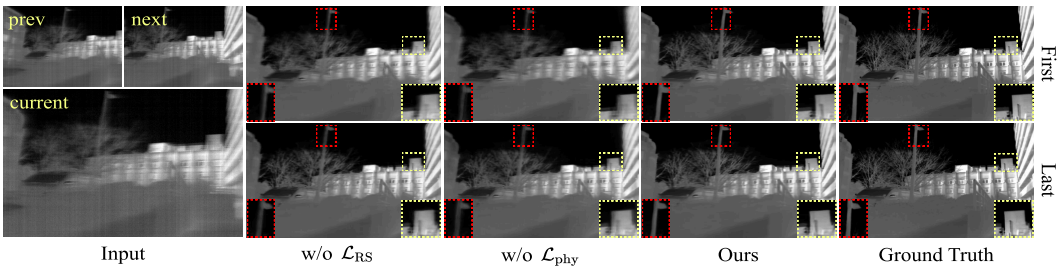


Figure 6: Visual comparison for the ablation of reconstruction constraints.

5.2 Ablation Study

We conduct ablation studies on key components of the proposed framework, including geometric rectification, forward imaging consistency, temporal alignment, and time-conditioned latent reconstruction. All variants are evaluated using the same protocol as the main comparison. The latter two studies are provided in the appendix.

Table 2: Quantitative results of the ablation study.

Method	*All Query Times*			*Final Query*		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o Motion Key	28.3224	0.9259	0.1180	28.0027	0.9191	0.1192
Implicit Rect.	28.0013	0.9298	0.1187	27.9581	0.9279	0.1217
w/o Row Time	28.1109	0.9249	0.1130	28.2150	0.9255	0.1128
w/o Fusion	28.2168	0.9280	0.0995	28.1173	0.9250	0.1014
w/o $\mathcal{L}_{\mathrm{RS}}$	27.3169	0.9224	0.1286	27.1230	0.9181	0.1162
w/o $\mathcal{L}_{\mathrm{phy}}$	27.5413	0.9194	0.1390	27.3463	0.9149	0.1293
Ours	28.7428	0.9334	0.0991	28.7565	0.9326	0.1006

Geometric rectification. We ablate four key components in the rectification branch. *w/o Motion Key* removes the motion augmentation from the attention key while preserving the remaining rectification pipeline, whereas *w/o Row Time* removes the explicit row-dependent temporal condition. *w/o Fusion* omits the final fusion block, while *Implicit Rect.* removes explicit distortion-flow estimation and geometric resampling, forcing direct GS prediction. As shown in Fig. 5, *w/o Motion Key* and *w/o Row Time* retain noticeable RS displacement in the yellow region. *w/o Fusion* introduces more visible artifacts at the first queried state, whereas *Implicit Rect.* partially corrects the distortion but degrades the window structure in the red region. The quantitative results in Table 2 consistently favor the full model.

Forward Imaging Consistency. We separately remove the direct RS-state supervision and the physical reblur consistency constraint, denoted as *w/o* $\mathcal{L}_{\mathrm{RS}}$ and *w/o* $\mathcal{L}_{\mathrm{phy}}$. *w/o* $\mathcal{L}_{\mathrm{RS}}$ removes the

direct constraint on individual reconstructed states, while *w/o* $\mathcal{L}_{\text{phy}}$ removes their joint consistency with the thermal image formation process. As shown in Fig. 6, the camera-motion scene is particularly challenging around the flagpole and rooftop regions, where all variants exhibit some residual distortion. Nevertheless, either ablation introduces more visible artifacts and degrades the corresponding edge and line structures. The full model produces cleaner reconstruction, consistent with the quantitative results in Table 2, supporting the complementarity of the two constraints.

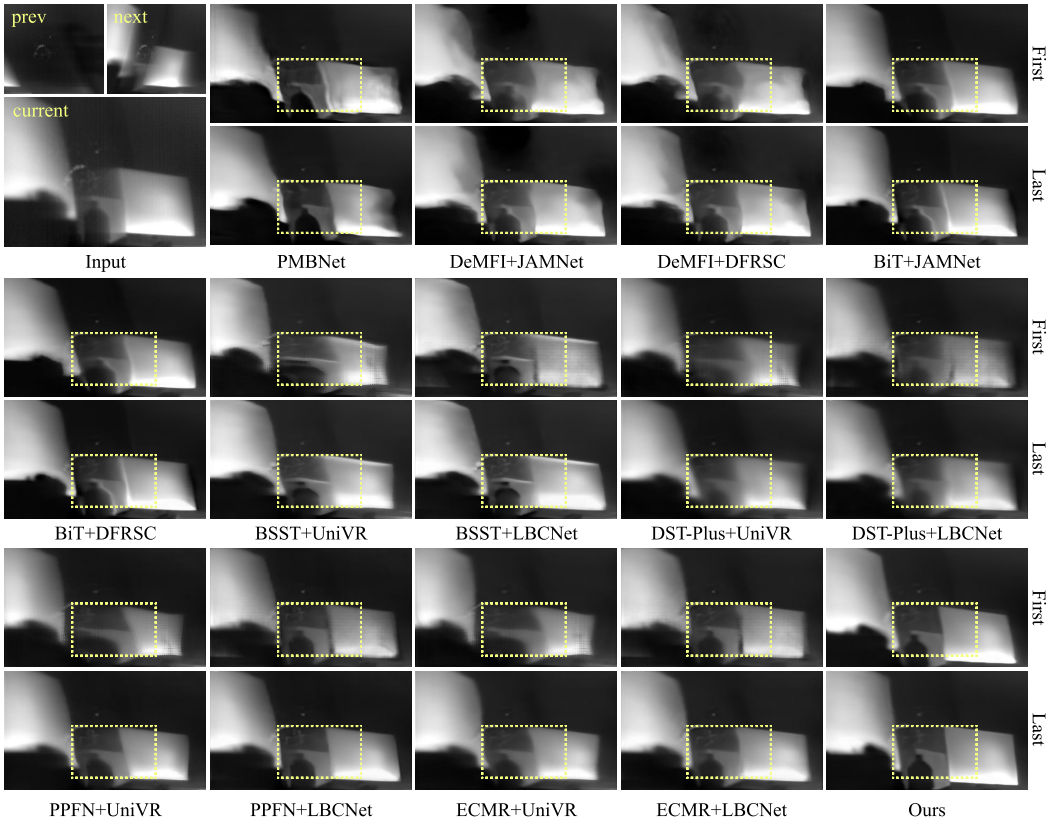


Figure 7: Visual comparison on the real-world dataset.

5.3 Demo on the Real-world Dataset

We qualitatively evaluate the proposed method on the real-world benchmark, for which sharp GS ground truth is unavailable. Fig. 7 shows results from the VIM384GS-USL camera. The indoor scene contains a bright computer screen, where camera motion induces severe thermal blur and rolling-shutter distortion. Most competing methods fail to simultaneously recover the bottle details and rectify the distorted screen, whereas our method reconstructs a clearer bottle, better restores the screen's vertical structure, and preserves more consistent temporal evolution across the reconstructed states. These results demonstrate the effectiveness of our method on real uncooled infrared observations. Additional HIKMICRO HM-TM06-LF/ANT comparisons are provided in the appendix, while video results from both real-world sensors are included in the supplementary material.

6 Conclusion

We presented a framework for thermal sequence reconstruction that jointly models microbolometer thermal persistence and rolling-shutter readout. Experiments demonstrate improved reconstruction performance and cross-sensor generalization on real uncooled infrared data.

AI USE STATEMENT

Generative AI tools were used for language polishing and to provide feedback on the presentation and experimental design of the manuscript. All research decisions and AI-assisted content were independently reviewed and verified by the authors, who take full responsibility for the final paper.

ETHICS STATEMENT

The real-world infrared data include human participants who were informed about the data collection procedure and provided consent for recording. The captured infrared imagery does not contain readily identifiable facial or personal details. The data were collected solely for research purposes, and we are not aware of any additional ethical concerns associated with the proposed methodology or experimental protocol.

REPRODUCIBILITY STATEMENT

We provide the formulation and model design in Secs. 3.1–3.4, the dataset construction and experimental settings in Secs. 4 and 5, and additional implementation details in Appendix A, including the synthetic data generation process, Wavelet State-Space Block, and motion-guided geometric rectification. The appendix also reports additional ablations, sequence visualizations, runtime comparisons, and real-world results.

REFERENCES

Mingdeng Cao, Zhihang Zhong, Jiahao Wang, Yinqiang Zheng, and Yujiu Yang. Learning adaptive warping for realworld rolling shutter correction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17764–17772. IEEE, 2022.

Mingdeng Cao, Sidi Yang, Yujiu Yang, and Yinqiang Zheng. Rolling shutter correction with intermediate distortion flow estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25338–25347. IEEE, 2024.

Spencer Carmichael, Manohar Bhat, Mani Ramanagopal, Austin Buchan, Ram Vasudevan, and Katherine A Skinner. Trnerf: restoring blurry, rolling shutter, and noisy thermal images with neural radiance fields. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7980–7990. IEEE, 2025.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024.

Bin Fan, Yuxin Mao, Yuchao Dai, Zhexiong Wan, and Qi Liu. Joint appearance and motion learning for efficient rolling shutter correction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5671–5681. IEEE, 2023.

Bin Fan, Yuchao Dai, and Hongdong Li. Learning bilateral cost volume for rolling shutter temporal super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3862–3879, 2024a.

Bin Fan, Zhexiong Wan, Boxin Shi, Chao Xu, and Yuchao Dai. Unified video reconstruction for rolling shutter and global shutter cameras. *IEEE Transactions on Image Processing*, 33:6821–6835, 2024b.

Elena Garcia, Maria Antonia Jimenez, Pablo Gonzalez De Santos, and Manuel Armada. The evolution of robotics research. *IEEE Robotics & Automation Magazine*, 14(1):90–103, 2007.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1777–1786. Ieee, 2020.

Muhammad Hasanujjaman, Mostafa Zaman Chowdhury, and Yeong Min Jang. Sensor fusion in autonomous vehicle with traffic surveillance camera system: detection, localization, and ai networking. *Sensors*, 23(6):3335, 2023.

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

Xiang Ji, Zhixiang Wang, Zhihang Zhong, and Yinqiang Zheng. Rethinking video frame interpolation from shutter mode induced degradation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12225–12234. IEEE, 2023.

Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In *european conference on computer vision*, pp. 663–678. Springer, 2022.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8104–8113. IEEE, 2019.

Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2070–2081. IEEE, 2023.

Kangwook Ko, Kyujin Shim, Kangil Lee, and Changick Kim. Large-scale benchmark for uncooled infrared image deblurring. *IEEE Sensors Journal*, 23(24):30119–30128, 2023.

Lingshun Kong, Jiangxin Dong, Jinhui Tang, Ming-Hsuan Yang, and Jinshan Pan. Efficient visual state space model for image deblurring. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12710–12719. IEEE, 2025.

Kangil Lee, Yuseok Ban, and Changick Kim. Motion blur kernel rendering using an inertial sensor: Interpreting the mechanism of a thermal detector. *Sensors*, 22(5):1893, 2022.

Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9822–9832. IEEE, 2023.

Pu Li, Huafeng Li, Yafei Zhang, Wen Wang, Neng Dong, and Jie Wen. Expandable, compressible, mineable: Open-world thermal image restoration. *arXiv preprint arXiv:2605.16967*, 2026.

Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13334–13343. PMLR, 17–23 Jul 2022.

Jinyuan Liu, Zihang Chen, Zhu Liu, Zhiying Jiang, Long Ma, Xin Fan, and Risheng Liu. Enhancing infrared vision: progressive prompt fusion network and benchmark. *Advances in Neural Information Processing Systems*, 38:96850–96875, 2026.

Muyao Niu, Zhuoxiao Li, Yifan Zhan, Huy H Nguyen, Isao Echizen, and Yinqiang Zheng. Physics-based adversarial attack on near-infrared human detector for nighttime surveillance camera systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8799–8807, 2023a.

Muyao Niu, Zhuoxiao Li, Zhihang Zhong, and Yinqiang Zheng. Visibility constrained wide-band illumination spectrum design for seeing-in-the-dark. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13976–13985. IEEE, 2023b.

Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. Nir-assisted video enhancement via unpaired 24-hour data. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10744–10754. IEEE, 2023c.

Muyao Niu, Tong Chen, Yifan Zhan, Zhuoxiao Li, Xiang Ji, and Yinqiang Zheng. Rs-nerf: Neural radiance fields from rolling shutter images. In *European Conference on Computer Vision*, pp. 163–180. Springer, 2024.

Muyao Niu, Mingze Ma, Yifan Zhan, Qingtian Zhu, Zhihang Zhong, Wei Guo, Chang Wen Chen, and Yinqiang Zheng. Toward robust and 3d-aware rgb-nir imaging in the dark. *arXiv preprint arXiv:2607.29684*, 2026a.

Muyao Niu, Yifan Zhan, Qingtian Zhu, Zhuoxiao Li, Wei Wang, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Motion-aware animatable gaussian avatars deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40140–40151, 2026b.

Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pp. 198–215. Springer, 2022.

Beata Oswald-Tranta, Mario Sorger, and Paul O'Leary. Motion deblurring of infrared images from a microbolometer camera. *Infrared physics & technology*, 53(4):274–279, 2010.

Jinshan Pan, Long Sun, Boming Xu, Jiangxin Dong, and Jinhui Tang. Learning efficient deep discriminative spatial and temporal networks for video deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5708–5722, 2025.

Delin Qu, Yizhen Lao, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Towards nonlinear-motion-aware and occlusion-robust rolling shutter correction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10646–10654. IEEE, 2023.

Manikandasriram Srinivasan Ramanagopal, Zixu Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. Pixel-wise motion deblurring of thermal videos. *arXiv preprint arXiv:2006.04973*, 2020.

Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5113–5122. IEEE, 2020.

Yang Tang, Chaoqiang Zhao, Jianrui Wang, Chongzhen Zhang, Qiyu Sun, Wei Xing Zheng, Wenli Du, Feng Qian, and Juergen Kurths. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 9604–9624, 2022.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Nisha Varghese, Mahesh Mohan MR, and AN Rajagopalan. Fast motion-deblurring of ir images. *IEEE Signal Processing Letters*, 29:459–463, 2022.

Martin Veit, Andrea Lucherini, Steven Verstockt, and Bart Merci. Video enhancement for increased spatio-temporal resolution in thermal videos: demonstration on a pool fire. *Fire Safety Journal*, pp. 104758, 2026.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17662–17672. IEEE, 2022.

Weilong Yan, Robby T Tan, Bing Zeng, and Shuaicheng Liu. Deep homography mixture for single image rolling shutter correction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9843. IEEE, 2023.

Zhanglei Yang, Haipeng Li, Shen Cheng, Mingbo Hong, Bing Zeng, and Shuaicheng Liu. Multi-frame rolling shutter correction with diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):10480–10492, 2025.

Shi Yi, Li Li, Xi Liu, Junjie Li, and Ling Chen. Hctirdeblur: A hybrid convolution-transformer network for single infrared image deblurring. *Infrared Physics & Technology*, 131:104640, 2023.

Chongchong Yu, Tianxin Bai, Wenbin Feng, Tong Zheng, and Tao Xie. Cfst: a cross-modal style transfer and flow-guided sparse transformer method for infrared video deblurring. *Infrared Physics & Technology*, pp. 106775, 2026.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-aware spatio-temporal sparse transformer for video deblurring. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3616–3626. IEEE, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European conference on computer vision*, pp. 191–207. Springer, 2020.

Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5713–5723. IEEE, 2023.

Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2482–2491, 2019.

Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4546–4555. IEEE, 2019.

A APPENDIX

Details for the Temporal Confidence-Controlled Fusion Module. Let F_c denote the center feature and F_n an aligned neighboring feature. TCCF preserves F_c as an identity path and injects each neighbor only through a gated residual:

$$F = F_c + s_{\text{prev}} G_{\text{prev}} \odot R_{\text{prev}} + s_{\text{next}} G_{\text{next}} \odot R_{\text{next}},$$

followed by a 3×3–GELU–3×3 residual refinement. For each neighbor, the residual branch is implemented as two 3×3 convolutions with GELU, while the gate is computed from the concatenation of the center feature, aligned neighboring feature, their channel-averaged absolute feature difference, and the gradient magnitude of the center degraded image. The gate branch also consists of two 3×3 convolutions with GELU, followed by a sigmoid. The previous and next branches use independent parameters. We initialize the last gate convolution with zero weights and bias -2, and initialize s_{prev} and s_{next} to 10^{-3}, such that fusion starts close to the center identity path and gradually introduces aligned neighboring information during training.

A.1 DETAILS FOR THE PROPOSED SYNTHETIC DATASET

Sharp infrared acquisition. The synthetic benchmark is constructed from high-frame-rate sharp infrared sequences captured using a cooled FLIR X6981-HS SLS-PH camera. It employs a 640×512 cooled SLS detector with a 25 μm pixel pitch, 14-bit dynamic range, a typical NETD of 40 mK, and snapshot readout. We operate the camera at 1,000 fps under full resolution to densely sample scene motion. The cooled high-speed detector suppresses thermal-response-induced trailing, while snapshot acquisition avoids row-dependent readout distortion, providing sharp GS source sequences for subsequent synthetic degradation.

Temporal densification. Although the source sequences are captured at 1,000 fps, their temporal spacing of 1 ms is still substantially coarser than the row-wise sampling interval required for RS synthesis. With a full-frame readout duration of $T_r = 32$ ms over 512 rows, adjacent rows correspond to a temporal difference of approximately 62.6 μs. Directly sampling the 1,000-fps sequence would therefore assign multiple neighboring rows to the same discrete source state, introducing artificial temporal quantization into the synthesized RS geometry. We consequently interpolate the source sequences by $8\times$ using FLAVR (Kalluri et al., 2023; Veit et al., 2026), reducing the temporal spacing to 0.125 ms and providing a substantially denser motion trajectory for row-dependent sampling.

Coupled thermal and rolling-shutter synthesis. We generate degraded observations according to the image formation model in Sec. 3.1. The RS readout duration is set to $T_r = 32$ ms, while the thermal response is modeled by a first-order exponential kernel with $\tau_{\text{th}} = 8$ ms. We truncate the response history at $T_m = 5\tau_{\text{th}} = 40$ ms, which preserves more than 99% of the exponential response energy. For numerical synthesis, the thermal history is discretized with an interval of $\Delta t = 2$ ms, resulting in 21 temporal taps. The normalized discrete response weights are given by $w_k = \exp(-k\Delta t/\tau_{\text{th}}) / \sum_j \exp(-j\Delta t/\tau_{\text{th}})$. For each image row, sharp scene states are sampled according to its row-dependent readout time and the corresponding temporal offsets, and are then accumulated using these response weights. Note that the $8\times$ temporal interpolation and the 2-ms thermal discretization serve different purposes: the former provides dense latent states for row-wise geometric sampling, whereas the latter approximates the temporal integration of the thermal response.

Supervision targets. For each synthesized observation, we additionally construct sharp RS and GS references from the same temporally densified sequence. The sharp RS target follows the row-dependent sampling time without thermal integration, thereby isolating the geometric distortion caused by RS acquisition. The corresponding GS target samples all rows at the queried time without row-dependent offsets. The nominal GS reference time is aligned with the center-row readout time of the corresponding RS observation, ensuring temporal correspondence between the two representations. These paired targets provide supervision for temporally resolved reconstruction and subsequent RS-to-GS geometric rectification.

Noise and dataset construction. After the coupled thermal-RS degradation is synthesized, sensor noise is added following the procedure of Liu et al. (2026) to approximate the degradation of real uncooled infrared observations. The resulting benchmark contains 9,015 training observations from

28 scenes and 1,038 test observations from 4 scenes. The split is strictly scene-disjoint, with all observations from each source sequence assigned exclusively to either the training or test set.

A.2 Details of Wavelet State-Space Block

Due to the non-instantaneous thermal response of microbolometer detectors, radiation received over an extended temporal history is accumulated into the current observation. Under scene or camera motion, this temporal accumulation is mapped into spatially distributed blur along motion trajectories, producing spatially varying degradation and weakening transient structures. Recovering a queried sharp state therefore requires contextual modeling beyond local neighborhoods.

State-space models provide an efficient way to establish such long-range dependencies through selective scanning with linear complexity. However, visual features are inherently two-dimensional, whereas the state-space recurrence operates on a one-dimensional sequence. Directly flattening a feature map using a fixed traversal order introduces a strong directional bias: spatially adjacent pixels may become distant in the scanned sequence, while information can only be propagated from previously visited positions. A straightforward solution is to perform multiple directional scans in parallel, but this substantially increases the computational cost. Instead, we employ a single selective scan in each WSSB and vary its effective traversal direction across consecutive blocks through lightweight geometric transformations.

Given an input feature $X_i \in \mathbb{R}^{H\times W\times C}$ of the i-th WSSB, we first transform its spatial arrangement as:

$$G_i = \mathcal{T}_i(X_i) = \begin{cases} \mathrm{Transpose}(X_i), & i \bmod 2 = 0, \\ \mathrm{Flip}_{HW}(X_i), & i \bmod 2 = 1, \end{cases} \tag{6}$$

where $\mathrm{Flip}_{HW}(\cdot)$ reverses both the horizontal and vertical dimensions. The transformed feature is subsequently processed by a single selective scan. In this way, consecutive WSSBs observe the feature map under different traversal orders without introducing parallel scanning branches.

Specifically, after layer normalization LN, the transformed feature is linearly projected and divided into two branches: $X_1, X_2 = \mathrm{Split}\left(\mathcal{P}_{\mathrm{in}}\left(\mathrm{LN}(G_i)\right)\right)$, where $\mathcal{P}_{\mathrm{in}}(\cdot)$ denotes the input projection. The first branch is locally enhanced before state-space modeling:

$$\widetilde{X}_1 = \mathrm{GELU}\left(\mathrm{DWConv}_{3\times 3}(X_1)\right). \tag{7}$$

The resulting two-dimensional feature is then serialized according to the current traversal order, yielding a sequence $\{\widetilde{x}_k\}_{k=1}^{HW}$. For each scanned position k, the selective state-space parameters are generated from the current feature content:

$$[\Delta_k, \mathbf{B}_k, \mathbf{C}_k] = \mathrm{DWConv}_{1D}\left(\mathcal{P}_s(\widetilde{x}_k)\right), \tag{8}$$

where Δ_k, $\mathbf{B}_k$, and $\mathbf{C}_k$ are input-dependent parameters generated from the current feature. Together with the learnable state transition and skip parameters $\mathbf{A}$ and $\mathbf{D}$, they parameterize the subsequent selective state-space recurrence. Using zero-order-hold discretization, the continuous transition parameters are converted to:

$$\overline{\mathbf{A}}_k = \exp(\Delta_k \mathbf{A}), \tag{9}$$

$$\overline{\mathbf{B}}_k = (\Delta_k \mathbf{A})^{-1}\left(\exp(\Delta_k \mathbf{A}) - \mathbf{I}\right)\Delta_k \mathbf{B}_k. \tag{10}$$

The hidden state is then recursively propagated along the scanned sequence:

$$\mathbf{h}_k = \overline{\mathbf{A}}_k \mathbf{h}_{k-1} + \overline{\mathbf{B}}_k \widetilde{x}_k, \tag{11}$$

$$\mathbf{y}_k = \mathbf{C}_k \mathbf{h}_k + \mathbf{D}\widetilde{x}_k, \tag{12}$$

where $\mathbf{h}_k$ summarizes information accumulated from previously visited spatial positions and $\mathbf{y}_k$ denotes the corresponding scan output. Since the state-space parameters depend on the current feature content, the propagation can adaptively retain or suppress information according to the spatially varying degradation. In our implementation, the state dimension is set to $d_{\mathrm{state}} = 8$, while the inner feature dimension is expanded to $2C$. After the selective scan, the output sequence is reshaped back to its current two-dimensional layout and modulated by the second projected branch:

$$S_i = \mathcal{P}_{\mathrm{out}}\left[\mathrm{LN}(Y_i) \odot \mathrm{GELU}(X_2)\right], \tag{13}$$

where Y_i denotes the reshaped state-space output, $\odot$ is element-wise multiplication, and $\mathcal{P}_{\text{out}}(\cdot)$ projects the feature back to C channels. The scan branch is finally incorporated through a residual connection:

$$X_i^{\text{scan}} = G_i + S_i. \tag{14}$$

A single selective scan imposes an inherent directional dependency, as each state aggregates information only from previously visited positions in the serialized sequence. The alternating transpose and flip operations modify this traversal order, thereby exposing the recurrent state propagation to different spatial orientations across consecutive WSSBs.

Notably, we do not immediately apply the inverse geometric transformation after an individual scan. The subsequent feed-forward refinement is performed in the current transformed coordinate system, and the following WSSBs continue to alter the orientation. The original spatial arrangement is recovered over a complete transformation cycle. This design avoids redundant transformations while enabling non-local information to propagate through different spatial traversal paths. For wavelet refinement, we employ Haar-initialized learnable analysis and synthesis filters, regularized by standard perfect-reconstruction and alias-cancellation constraints on the filter coefficients.

A.3 Details of Motion-Guided Geometric Rectification

The time-conditioned reconstruction module produces a sharp feature Z_t^{RS} at the queried time t, but this feature remains defined on the center RS sampling grid. We therefore explicitly estimate the geometric mapping from the target GS grid to the observed RS geometry and use it to rectify Z_t^{RS}. To preserve the native row-dependent geometry, the rectification branch reuses the frame-wise features extracted by the shared stem before temporal alignment, rather than the center-aligned feature F_{agg}.

Geometry and motion representation. The pre-alignment features of the previous, center, and next frames are processed by a shared convolutional pyramid. Although five scales are constructed, the rectification head uses only the full-, $1/4$-, and $1/16$-resolution features. At each scale, we additionally reuse the four bidirectional motion fields estimated by the frozen BFE between the previous–center and center–next frames. For a queried time t, the normalized row-time offset map is defined as $T_t(y) = \bar{t}_r(y) - t$, where $\bar{t}_r(y)$ denotes the row readout time $t_r(y)$, normalized to the same temporal coordinate as t. The row coordinates are constructed on the original sensor height before cropping. The four normalized motion fields and T_t are concatenated into a 9-channel descriptor and encoded into a 48-channel motion representation M^l.

Motion-guided coarse correspondence. Initial correspondence is estimated at the $1/16$ resolution. The three geometry features, motion representation, and row-time map are first fused to construct a query seed Q. Since thermal blur can weaken appearance cues, the independently estimated inter-frame motion is further introduced into the matching key. Specifically, let K_i^a denote the appearance key of frame i and Φ_i^m its encoded motion context. We use

$$K_i = K_i^a + \alpha_{\max} \tanh(\alpha_{\text{raw}}) \mathcal{P}_m(\Phi_i^m), \qquad \alpha_{\max} = 0.1, \tag{15}$$

where α_{raw} is a learnable scalar initialized to zero, while $\alpha_{\max}$ is a fixed constant that bounds the magnitude of the motion augmentation. The query and value branches remain unchanged. For the center frame, the motion context is obtained by averaging the motion representations from the two neighboring frame pairs.

Following the global correspondence formulation, the value contains both the geometry feature and normalized spatial coordinates, i.e., $V_i = [F_i^4, P]$. The attention output $A_i = \text{Softmax}(QK_i^\top/\sqrt{d})$ therefore provides both matched features and corresponding coordinates. The resulting displacement $\Delta P_i = \widetilde{P}_i - P$ is used as an intermediate correspondence cue rather than directly as the distortion flow. The correspondence features of all three frames are fused with M^4, after which a 3×3 flow head predicts three coarse GS-to-RS distortion fields $\{\mathbf{u}_{-1}^4, \mathbf{u}_0^4, \mathbf{u}_{+1}^4\}$.

Coarse-to-fine flow refinement. The coarse correspondence is estimated at $l = 4$, corresponding to the $1/16$ resolution. We then progressively refine the distortion fields at the remaining used levels $l \in \{2, 0\}$, corresponding to the $1/4$ and full resolutions, respectively. At each such level, let $\overline{\mathbf{u}}_i^l$

and $\overline{G}^l$ denote the distortion field and fused feature from the next coarser used scale, bilinearly upsampled to the current resolution, with the flow displacement magnitudes scaled accordingly. The fused feature is additionally projected to the current feature space.

The upsampled flow $\overline{\mathbf{u}}_i^l$ is first used to backward warp the corresponding current-scale geometry feature as

$$\widetilde{F}_i^l = \mathcal{W}(F_i^l, \overline{\mathbf{u}}_i^l),$$

where $\mathcal{W}(\cdot)$ denotes bilinear backward warping. The warped geometry features are then combined with the upsampled feature and flow estimates, together with the motion and temporal conditions:

$$G^l = \mathcal{R}^l\left([\widetilde{F}_{-1}^l, \widetilde{F}_0^l, \widetilde{F}_{+1}^l, \overline{G}^l, \overline{\mathbf{u}}_{-1}^l, \overline{\mathbf{u}}_0^l, \overline{\mathbf{u}}_{+1}^l, M^l, T_t^l]\right).$$

Here, $\overline{G}^l$ denotes the upsampled and projected fused feature from the next coarser used scale, M^l is the current-scale motion representation, T_t^l is the row-time offset map, and $\mathcal{R}^l$ consists of a 3×3 projection followed by two residual convolutional blocks.

The resulting feature G^l is finally passed through a 3×3 flow head to directly predict three distortion fields for the previous, center, and next frames:

$$\{\mathbf{u}_{-1}^l, \mathbf{u}_0^l, \mathbf{u}_{+1}^l\} = \mathcal{H}_{\text{flow}}^l(G^l).$$

Unlike residual flow refinement, the complete distortion fields are re-estimated at each used scale. This procedure is successively applied at the $1/4$ and full resolutions.

Explicit rectification and feature fusion. The reconstructed feature Z_t^{RS} is defined on the center RS grid. Therefore, we use the final center distortion field to obtain the rectified representation as $\widetilde{Z}_t^{GS} = \mathcal{W}(Z_t^{RS}, \mathbf{u}_0^0)$, where $\mathcal{W}(\cdot)$ denotes the bilinear backward warping operation defined above. Under large geometric displacement, some target locations may fall outside the spatial support of the center RS feature, leaving invalid regions after warping. The neighboring frames provide complementary information for these regions. However, their features are still extracted from degraded observations affected by thermal blur and sensor noise. To avoid unnecessarily perturbing regions that are already reliably reconstructed from the center frame, we restrict the neighboring contributions to the out-of-bounds regions of the center warp. Specifically, the previous and next geometry features are warped using $\mathbf{u}_{-1}^0$ and $\mathbf{u}_{+1}^0$, respectively, and only their responses within these unsupported regions are retained, yielding F_{-1}^{fill} and F_{+1}^{fill}. Finally, the center-warped temporal feature, the two neighboring completion features, and the geometry-aware feature G^0 from the final full-resolution refinement stage are jointly fused:

$$\Delta Z_t^{\text{GS}} = \mathcal{C}_{1\times1}\left([\widetilde{Z}_t^{\text{GS}}, F_{-1}^{\text{fill}}, F_{+1}^{\text{fill}}, G^0]\right), \tag{16}$$

where $\mathcal{C}_{1\times1}$ is a zero-initialized 1×1 convolution. Here, G^0 carries the geometry-aware context produced by the final flow-refinement stage. The fused correction ΔZ_t^{GS} is added to $\widetilde{Z}_t^{GS}$ and further refined by two residual convolutional blocks, followed by a final 3×3 convolution to produce the final GS image.

B Additional Ablation Study

Time-conditioned latent representation. The WSSB contains two functionally distinct stages: selective scanning for non-local information propagation and post-scan feature refinement for restoration. We therefore ablate both the spatial traversal strategy and the refinement design. Specifically, *Fixed Scan* removes the alternating flip and transpose transformations across successive WSSBs, such that all blocks follow the same traversal orientation. For post-scan refinement, we replace our design with several representative restoration alternatives while keeping the proposed scan strategy unchanged. *Fourier Refine.* (Kong et al., 2025) combines gated spatial mixing with patch-wise Fourier processing, providing a direct comparison between Fourier- and wavelet-domain refinement. *Gated Conv.* (Zamir et al., 2022) retains gated local transformation without explicit frequency decomposition, while *Local Conv.* (Wang et al., 2022) employs conventional local convolutional processing without gating or explicit frequency refinement. Together, these variants examine the

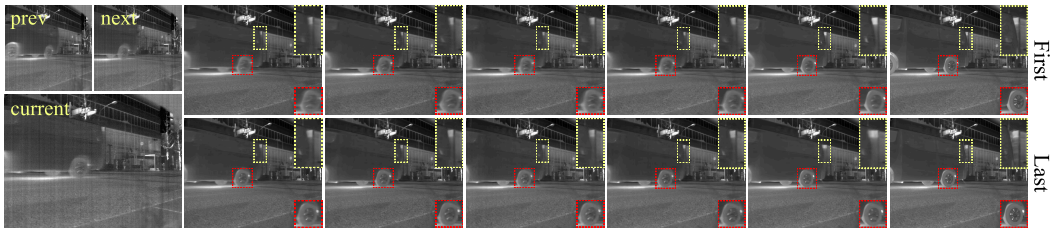


Figure 8: Visual comparison for the ablation of time-conditioned RS reconstruction.

Table 3: Quantitative results on the synthetic dataset.

Method	All Query Times			Final Query		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fixed Scan	27.0893	0.9185	0.1509	26.8345	0.9143	0.1401
Fourier Refine.	27.3163	0.9218	0.1219	26.9491	0.9175	0.1177
Gated Conv.	27.4765	0.9215	0.1323	27.3225	0.9160	0.1240
Local Conv.	27.6419	0.9170	0.1322	27.2660	0.9125	0.1224
w/o TCCF	27.7478	0.9206	0.1458	27.4056	0.9153	0.1370
w/o Flow	27.6449	0.9221	0.1297	27.3791	0.9177	0.1216
Implicit Prop.	27.6700	0.9217	0.1378	27.4274	0.9163	0.1321
PWC-Net Flow	28.5442	0.9329	0.1000	28.3551	0.9304	0.0992
Ours	28.7428	0.9334	0.0991	28.7565	0.9326	0.1006

contributions of traversal diversity, frequency-aware refinement, and gated local transformation in the WSSB.

As shown in Fig. 8, the moving vehicle contains severe blur and RS deformation around the wheel highlighted in red. The ablated variants leave more noticeable blur or structural distortion, whereas the full WSSB recovers a sharper and more coherent wheel contour. In particular, *Fixed Scan* shows similar degradation, indicating the benefit of varying the effective scan orientation across successive blocks. In the yellow region with RS-induced occlusion, the full model also restores more complete structures than the alternatives. The quantitative results in Table 3 consistently favor the complete WSSB design.

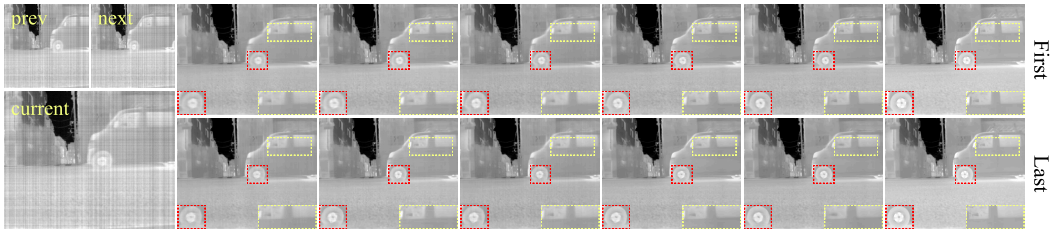


Figure 9: Visual comparison for the ablation of temporal alignment.

Temporal alignment and fusion. A potential concern in our temporal aggregation is the reliability of optical flow estimated from blurred infrared observations. We therefore design a set of controlled variants to examine both the necessity of explicit motion guidance and the sensitivity of our model to imperfect or estimator-dependent flow. First, *w/o Flow* removes explicit flow warping and directly aggregates the multi-frame features, evaluating whether motion-guided alignment remains beneficial despite the domain gap. *w/o TCCF* retains the estimated flow but removes the confidence-controlled fusion mechanism, testing whether the center-dominant gated fusion is important for suppressing unreliable aligned features. *Implicit Prop.* (Li et al., 2023) replaces explicit flow-based alignment with implicit feature propagation, providing an alternative that does not rely on direct geometric warping. Finally, *PWC-Net Flow* replaces the default RAFT estimator with PWC-Net while keeping the remaining framework unchanged, allowing us to assess the dependence on a particular optical-flow model.

As shown in Fig. 9, *w/o Flow* results in less accurate recovery of the wheel structure and fine details, while *w/o TCCF* introduces more visible degradation when unreliable aligned features are directly incorporated. *Implicit Prop.* also retains noticeable RS distortion around the wheel in the red region, indicating that explicit motion-guided alignment provides more effective temporal correspondence for this task. In contrast, *PWC-Net Flow* produces visual results very close to those obtained with RAFT, with similarly clean structures in the highlighted regions. The quantitative results in Table 3 show the same trend: explicit flow guidance and confidence-controlled fusion are both beneficial, while replacing RAFT with PWC-Net causes only a marginal change in performance. These results indicate that the proposed temporal aggregation benefits from optical-flow guidance without being tightly coupled to a specific flow estimator, and remains effective under the imperfect motion estimation encountered in blurred infrared imagery.

C Visualization of Sequence Reconstruction

The main paper presents only representative first and last query due to space limitations. We further visualize six reconstructed states on the synthetic benchmark to examine the temporal evolution recovered by our model. As shown in Fig. 10, the scene contains two vehicles moving in opposite directions. From the first to the last reconstructed state, the two vehicles progressively move apart, while their structures remain sharp and geometrically coherent throughout the sequence. The consistent evolution of their relative positions demonstrates that our model recovers temporally ordered scene states rather than independently restoring individual frames, supporting its ability to reconstruct the underlying motion from thermal blur.

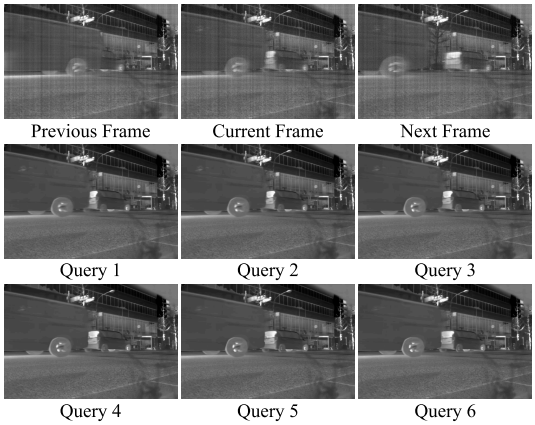


Figure 10: Sequence reconstruction visualization on the proposed synthetic dataset.

D Inference Efficiency Comparison

We further report the average inference time on the proposed synthetic test set to provide a reference for computational efficiency. To avoid conflating runtime with the number of reconstructed temporal states, all methods are evaluated for generating a single target frame rather than the full six-frame sequence. As shown in Table 4, our method requires 0.5570 s per frame. Although it is not the fastest approach, it remains faster than most of the compared restoration pipelines, including the DeMFI- and BiT-based combinations as well as several cascaded deblurring–RSC variants. The fastest methods achieve lower latency, but the proposed framework maintains a moderate inference cost despite jointly performing temporally resolved restoration and geometric rectification. Since the compared methods differ substantially in architecture and implementation, these timings are intended primarily as an efficiency reference rather than a strict complexity ranking.

Table 4: Average inference time on the proposed synthetic dataset.

Method	*Average Time (s)*
PMBNet	0.6554
DeMFI+JAMNet	2.4751
DeMFI+DFRSC	2.6428
BiT+JAMNet	1.2968
BiT+DFRSC	1.4613
BSST+UniVR	1.3591
BSST+LBCNet	1.7390
DST-Plus+UniVR	0.2098
DST-Plus+LBCNet	0.5906
PPFN+UniVR	0.4899
PPFN+LBCNet	0.8726
ECMR+UniVR	0.1566
ECMR+LBCNet	0.5362
Ours	0.5570

E Additional Demo on the Real-world Dataset

We further evaluate the proposed model on real sequences captured by the HIKMICRO HM-TM06-LF/ANT. As shown in Fig. 11, the fast camera motion produces severe thermal blur together with pronounced RS distortion around the building entrance. Most competing methods either leave substantial motion blur or produce noticeable geometric deformation, causing the gate structure to remain tilted or distorted. In contrast, our method recovers the dominant structure of the entrance more



Figure 11: Visual comparison on the real-world dataset.

clearly while better preserving the vertical alignment of the gate. Together with the results obtained on the Vision Sensing camera, these observations demonstrate favorable cross-sensor generalization of the proposed method.