

HATIR: Heat-Aware Diffusion for Turbulent Infrared Video Super-Resolution

Yang Zou¹, Xingyue Zhu², Kaiqi Han², Jun Ma², Xingyuan Li³, Zhiying Jiang⁴, Jinyuan Liu^{2*}

¹Northwestern Polytechnical University, Xi'an, China

²Dalian University of Technology, Dalian, China

³Zhejiang University, Hangzhou, China

⁴Dalian Maritime University, Dalian, China

archerv2@mail.nwpu.edu.cn, atlantis918@hotmail.com

Abstract

Infrared video has been of great interest in visual tasks under challenging environments, but often suffers from severe atmospheric turbulence and compression degradation. Existing video super-resolution (VSR) methods either neglect the inherent modality gap between infrared and visible images or fail to restore turbulence-induced distortions. Directly cascading turbulence mitigation (TM) algorithms with VSR methods leads to error propagation and accumulation due to the decoupled modeling of degradation between turbulence and resolution. We introduce **HATIR**, a **Heat-Aware Diffusion for Turbulent InfraRed Video Super-Resolution**, which injects heat-aware deformation priors into the diffusion sampling path to jointly model the inverse process of turbulent degradation and structural detail loss. Specifically, HATIR constructs a Phasor-Guided Flow Estimator, rooted in the physical principle that thermally active regions exhibit consistent phasor responses over time, enabling reliable turbulence-aware flow to guide the reverse diffusion process. To ensure the fidelity of structural recovery under nonuniform distortions, a Turbulence-Aware Decoder is proposed to selectively suppress unstable temporal cues and enhance edge-aware feature aggregation via turbulence gating and structure-aware attention. We built FLIR-IVSR, the first dataset for turbulent infrared VSR, comprising paired LR-HR sequences from a FLIR T1050sc camera (1024×768) spanning 640 diverse scenes with varying camera and object motion conditions. This encourages future research in infrared VSR.

Code — <https://github.com/JZ0606/HATIR>

Introduction

High-quality infrared (IR) video is critical for vision tasks in challenging environments, such as autonomous driving, surveillance, and object tracking (Liu et al. 2025b; Wang et al. 2025a). However, infrared imaging systems deployed in open atmospheric environments are highly susceptible to degradation caused by atmospheric turbulence. The formation of such turbulence is primarily attributed to the thermal and dynamic instability within the atmospheric boundary layer. Specifically, the temperature gradients between the

hot ground surface and the cooler upper atmosphere generate convective flows that lead to the emergence of turbulent eddies across multiple spatial and temporal scales, as shown in Figure 1. These turbulent eddies cause random fluctuations of the refractive index and thermal radiation in the turbulence medium, which bend the propagated wave, resulting in geometric distortions, thermal blur, and grayscale drift in the infrared imaging (Wang et al. 2023; Zou et al. 2026). Compared to visible light cameras, IR sensors are more susceptible to turbulence-induced distortions due to their longer wavelengths and sensitivity to thermal fluctuations (Liu et al. 2022; Li et al. 2024, 2025b; Liu et al. 2024). These real-world factors make the acquisition of high-quality IR video particularly challenging in practical scenarios.

Conventionally, sliding-window based VSR methods (Haris, Shakhnarovich, and Ukita 2019; Tian et al. 2020; Wang et al. 2019) reconstruct a high-resolution (HR) video by extracting features from a fixed number of adjacent frames within a short temporal window. Recurrent methods (Huang, Wang, and Wang 2015, 2017; Isobe et al. 2020; Sajjadi, Vemulapalli, and Brown 2018) propagate hidden features by capturing long-term temporal dependencies and exploiting motion continuity across frames. Recently, diffusion-based methods (Zhou et al. 2024a; Yang et al. 2024; Chen et al. 2024; Zhao et al. 2023) have demonstrated remarkable performance in generating high-fidelity and perceptually realistic video content. These approaches primarily focus on incorporating temporal consistency strategies into the diffusion framework.

Despite the remarkable progress of video super-resolution (VSR), existing approaches face two fundamental challenges when applied to infrared videos with turbulence: 1) **Modality gap**. Infrared images exhibit low texture contrast, weak structural boundaries, and thermal-dominated intensity patterns, deviating significantly from the assumptions underlying RGB-based VSR models (Liu et al. 2022, 2025a; Zou et al. 2024; Li et al. 2022, 2025a; Wang et al. 2025b). 2) **Turbulence ignorance**. Severe atmospheric turbulence introduces nonlinear geometric distortions and unstable thermal boundaries, which are not explicitly addressed by conventional VSR pipelines. While turbulence mitigation (TM) methods fail to recover structural details. Simply cascading TM with VSR models often causes **error propagation and accumulation** due to their decoupled nature. Given these

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

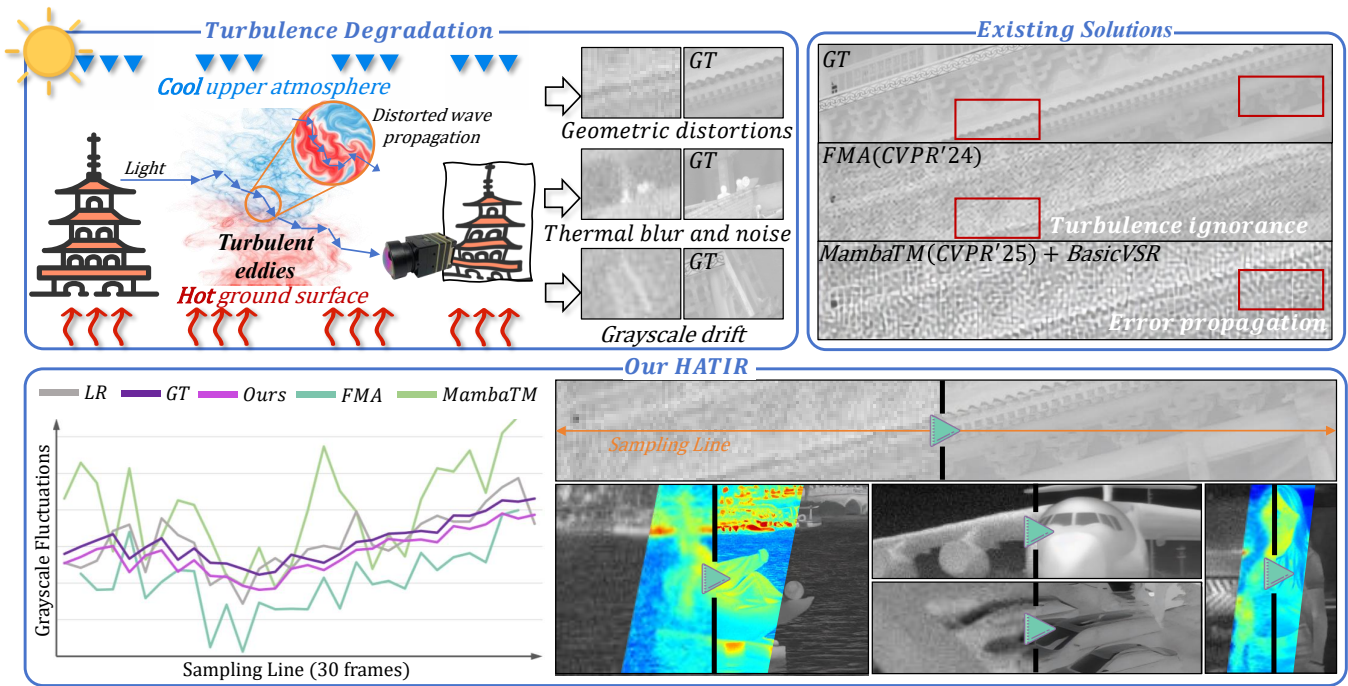


Figure 1: Infrared VSR performance under turbulence conditions evaluated by HATIR on the proposed FLIR-IVSR dataset. The graph illustrates grayscale fluctuations along the orange-marked sampling line over time (30 video frames).

challenges, we ask, “Is it possible to solve the turbulent infrared VSR through a unified inverse process?”

The answer is “Yes.” We propose HATIR, a Heat-Aware Diffusion framework for Turbulent InfraRed Video Super-Resolution, which injects physically grounded heat-aware deformation priors into the diffusion sampling path to jointly model the inverse process of turbulence degradation and structural detail loss. By unifying alignment and restoration in a single generative path, HATIR mitigates error amplification caused by misalignment and thermal blur, which conventional approaches often struggle with. Specifically, we propose Phasor-Guided Flow Estimator (PhasorFlow), enabling robust turbulence-aware motion guidance. Also, a Turbulence-Aware Decoder (TAD) is introduced to enhance structural fidelity under non-uniform distortions via turbulence-aware gating and structure-aware feature fusion. To benchmark this task, we construct the first dataset for turbulent infrared VSR, enabling evaluation under long-range infrared degradation. Our contribution can be summarized as follows:

- We introduce **HATIR**, a **H**eat-Aware Diffusion for **T**urbulent **I**nfra**R**ed Video Super-Resolution, which jointly models the degradation process of turbulent degradation and structural detail loss through physics-driven heat-aware deformation priors.
- We design a phasor-guided flow estimator, rooted in thermal consistency, to provide robust turbulence-aware guidance for reverse diffusion. A Turbulence-Aware Decoder is further introduced to enhance structural restoration by suppressing unstable temporal information and

reinforcing edge-aware feature aggregation.

- We built the first dataset for the turbulent infrared VSR task, FLIR-IVSR, comprising paired LR-HR sequences captured by a FLIR T1050sc camera at a resolution of 1024×768 . FLIR-IVSR spans 640 diverse scenes under varying camera and object motion conditions.

Related Work

Video Super-Resolution

Existing VSR methods can be broadly categorized into multiple-input single-output (MISO) and multiple-input multiple-output (MIMO) paradigms. MISO-based methods reconstruct the center frame from a fixed window of LR frames. This line of work includes filter-based approaches (Jo et al. 2018), alignment-based methods using deformable convolutions (Wang et al. 2019), and attention-based designs (Li et al. 2020). Recent extensions further integrate motion-aware modules (Youk, Oh, and Kim 2024), recurrent propagation (Chi et al. 2024), or G-buffer priors (Zheng et al. 2025) for enhanced temporal modeling and efficiency. MIMO-based methods jointly reconstruct multiple frames, allowing for consistent modeling across time. This includes transformer-based architectures (Liang et al. 2022) and diffusion-driven approaches (Yang et al. 2024; Zhou et al. 2024a), which incorporate motion priors into the generative process to improve fidelity and coherence.

Video Turbulence Mitigation

Traditional methods typically employ a three-stage pipeline comprising registration, fusion, and deblurring. Recent

learning-based methods address turbulence dynamics in an end-to-end manner. DATUM (Zhang et al. 2024) decouples alignment and content restoration across short sequences. MambaTM (Zhang et al. 2025) adopts state space models for efficient long-range temporal modeling. Turb-Seg-Res (Saha et al. 2024) separates motion-dominant regions for region-specific refinement. Nevertheless, these methods are designed for RGB videos and struggle in infrared domains due to weak textures and thermal blur. Moreover, they typically address turbulence alone, overlooking the resolution degradation that coexists in real infrared settings. This highlights the need for a unified solution to jointly mitigate turbulence and enhance resolution in infrared videos.

Method

Overview

As illustrated in Figure 2, the LR video is first encoded into a latent space via a VAE encoder. Then, guided by the proposed PhasorFlow, which captures the thermal dynamics of time-varying heat sources, the diffusion model iteratively refines the latent variables under turbulence-aware modulation. Finally, a Turbulence-Aware Decoder (TAD) reconstructs the HR frames by suppressing unreliable temporal cues and reinforcing edge structures.

Phasor-Guided Flow Estimator

To tackle turbulence-induced distortions and detail degradation in low-resolution infrared videos, we propose Phasor-Guided Flow Estimator (PhasorFlow), a heat-aware flow estimator that guides diffusion sampling with thermal priors as shown in Figure 3. While prior works leverage optical flow for inter-frame alignment (Yang et al. 2024; Liang et al. 2022; Zhou et al. 2024a; Wang et al. 2024), they often fail in turbulent infrared settings due to weak textures, ambiguous boundaries, and the stochastic nature of turbulence. PhasorFlow addresses these issues by introducing Frequency-Weighted Attention, guided by thermal phasor analysis, which measures the temporal consistency of thermal radiation in the frequency domain.

Specifically, we first extract shallow features $F^0 \in \mathbb{R}^{T \times H \times W \times C}$ and segment them into short clips. For each clip F_t^i , an initial flow $f_{t-1 \rightarrow t}^i$ is estimated via a pretrained flow network (Ranjan and Black 2017), and iteratively refined using the Phasor Mask and Frequency-Weighted Attention in a locally parallel, globally recurrent manner.

Phasor Mask To robustly identify thermally stable regions under turbulence, we calculate the Phasor Mask to assess the temporal frequency response of infrared sequences. This is based on the physical observation that heat-emitting regions exhibit stable temporal dynamics, while turbulence causes high-frequency, spatially varying perturbations.

Given a short infrared sequence $\mathbf{I} \in \mathbb{R}^{B \times T \times 1 \times H \times W}$, we first reshape it to $\mathbf{I}' \in \mathbb{C}^{B \times H \times W \times T}$ and compute the discrete Fourier transform (DFT) over the temporal dimension as $\hat{\mathbf{I}}(x) = \mathcal{F}_t(\mathbf{I}(x, :))$, $x \in \Omega$. We then extract the magnitude of the first harmonic (e.g., $\hat{\mathbf{I}}_1(x)$) as the primary frequency response by $M_{\text{phasor}}(x) = \left| \hat{\mathbf{I}}_1(x) \right|$. Finally, M_{phasor}

is normalized to $[0, 1]$ to serve as a soft mask:

$$M_{\text{phasor}}(x) = \sigma(\alpha \cdot (M_{\text{phasor}}(x) - \mu)), \quad (1)$$

where μ is the spatial mean and α is a scaling factor. This Phasor Mask emphasizes pixels with consistent temporal thermal signatures and is integrated into attention modulation and flow guidance to suppress unstable turbulent regions and preserve heat-sensitive structural information.

Frequency-weighted Attention Given the $(t-1)$ -th clip feature F_{t-1}^i from the i -th layer, our objective is to estimate the turbulence-mitigated flow $\hat{f}_{t-1 \rightarrow t}^{i, (1:N)}$ across the N frames in each clip. For each flow $\hat{f}_{t-1 \rightarrow t, n'}^{i, (n)}$ aligning frame n' in clip $t-1$ to frame n in clip t , we first compute a coarse optical flow $f_{t-1 \rightarrow t}^{i, (1:N)}$ using SpyNet (Ranjan and Black 2017), and obtain coarse aligned features via:

$$\bar{F}_{t-1}^{i, (1:N)} = \text{Warp}(F_{t-1}^i, M_{\text{phasor}, t-1 \rightarrow t}^{(1:N)} \circ f_{t-1 \rightarrow t}^{i, (1:N)}), \quad (2)$$

where M_{phasor} denotes the thermal stability prior from Phasor Mask. These coarse features are concatenated with the current frame and flow to predict flow residuals via a CNN:

$$\Delta f_{t-1 \rightarrow t}^{i, (1:N)} = \text{Conv}(\text{Concat}(\bar{F}_{t-1}^{i, (1:N)}, F_t^{i-1}, f_{t-1 \rightarrow t}^{i, (1:N)})). \quad (3)$$

We then update the flow through an averaged refinement across M predicted offsets:

$$f_{t-1 \rightarrow t, n'}^{i+1, (n)} = f_{t-1 \rightarrow t, n'}^{i, (n)} + \frac{1}{M} \sum_{m=1}^M \{\Delta f_{t-1 \rightarrow t, n'}^{i, (n)}\}_m, \quad (4)$$

where $\{\Delta f_{t-1 \rightarrow t, n'}^{i, (n)}\}_m$ denotes the m -th offset in total M predictions.

To enhance feature reliability during turbulence, we sample features via the updated flow and apply phasor-guided attention. Specifically, the attention queries, keys, and values are defined as $Q = F_{t, n}^{i-1} P_Q$, $K = \text{Sampling}(F_{t-1}^{i-1} P_K, f + \Delta f)$, and $V = \text{Sampling}(F_{t-1}^i P_V, f + \Delta f)$, where $f + \Delta f$ denotes the total motion offset. The Phasor Mask modulates attention weights as:

$$\hat{F}_{t-1}^{i, (n)} = (M_{\text{phasor}}^{(n)} \circ \mathcal{S}(QK^\top / \sqrt{C}))V + \text{MLP}(\hat{F}_{t-1}^{i, (n)}), \quad (5)$$

where \mathcal{S} denotes the SoftMax operation. In the final layer L , we recompute the offset using the refined feature $\hat{F}_{t-1}^{L, (1:N)}$ to update the final flow:

$$f_{t-1 \rightarrow t, n'}^* = f + \frac{1}{M} \sum_{m=1}^M \left[\overbrace{\mathcal{H}(\hat{F}_{t-1}^{L, (1:N)}, F_t^{L-1}, f_{t-1 \rightarrow t}^{L, (1:N)})}^{\Delta f_{t-1 \rightarrow t}^{L, (1:N)}} \right]_{n'}^{(m)}, \quad (6)$$

where f represents $f_{t-1 \rightarrow t, n'}^L$, $\mathcal{H}(\cdot)$ denotes a lightweight convolutional network.

Heat-aware Guidance To improve the stability and consistency of the denoising trajectory under turbulence, we inject a physics-informed guidance term derived from thermal

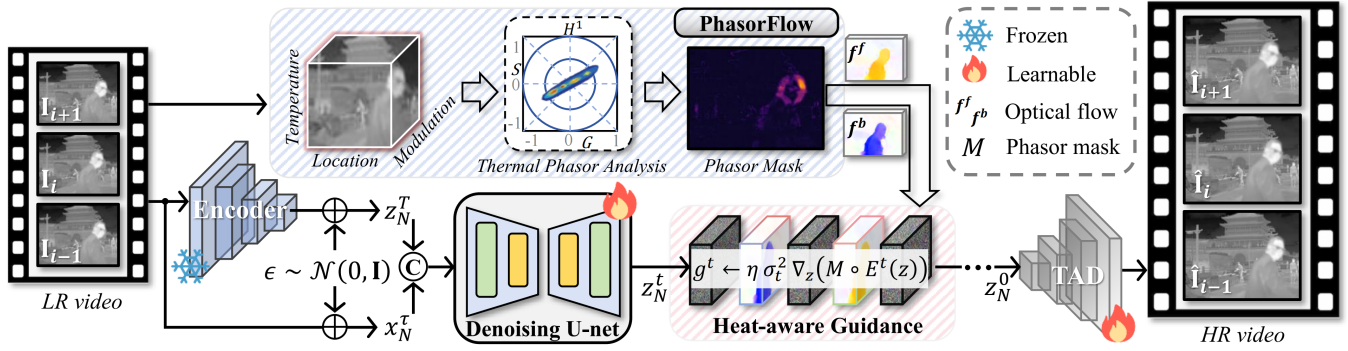


Figure 2: Given a low-resolution (LR) turbulent infrared video sequence $\mathbf{I}_{LR} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N\}$, HATIR reconstructs a high-resolution (HR) sequence $\mathbf{I}_{HR} = \{\hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2, \dots, \hat{\mathbf{I}}_N\}$ with suppressed turbulence distortions and enhanced temporal coherence. The proposed unified latent diffusion framework jointly addresses spatial degradation removal and inter-frame alignment for infrared videos under atmospheric turbulence.

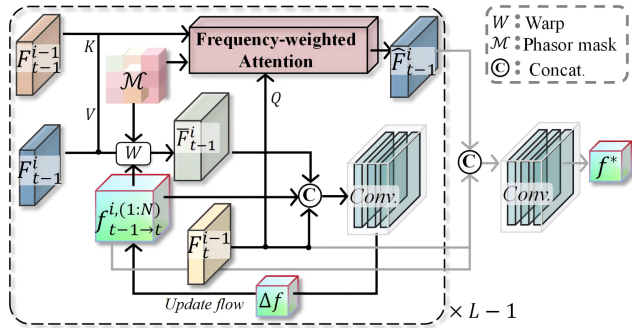


Figure 3: Overview of PhasorFlow.

motion priors. At each denoising step t , we first define the symmetric warping error between bidirectional flows:

$$E^t(z) = \sum_{i=1}^{N-1} \|\text{Warp}(z_i^t, f_{b,i}^*) - z_{i+1}^t\|_1 + \sum_{i=2}^N \|\text{Warp}(z_i^t, f_{f,i-1}^*) - z_{i-1}^t\|_1, \quad (7)$$

where $f_{f,i-1}^*$ and $f_{b,i}^*$ are the forward and backward flows estimated by PhasorFlow. To localize reliable temporal structures, we construct a heat-aware modulation mask M_{joint} by fusing an occlusion-aware mask and the normalized thermal Phasor Mask as $M_{\text{joint}} = M_{\text{occ}} \cdot M_{\text{phasor}}$, where M_{phasor} denotes the Phasor Mask.

The final heat-aware guidance term is defined as $g^t = \eta \sigma_t^2 \nabla_z (M_{\text{joint}} \circ E^t(z))$, where σ_t^2 is the noise variance at step t , and η modulates the influence of the guidance. The denoising step is then adjusted as:

$$\hat{z}^t = z^{t+1} - \sigma_t^2 \epsilon_\phi(z^{t+1}, t) - g^t, \quad (8)$$

where ϵ_ϕ denotes the noise prediction network of the diffusion model. This guidance steers the sampling trajectory toward temporally coherent and thermally stable representations, which are subsequently decoded by the Turbulence-Aware Decoder (TAD).

Turbulence-Aware Decoder

IR images typically exhibit weak textures, blurred thermal boundaries, and reduced structural saliency compared to visible images. These properties, compounded by atmospheric turbulence, result in alignment errors and unreliable motion estimation. Also, enforcing strict temporal consistency in turbulence-distorted regions may introduce erroneous corrections. Given those issues, we propose the Turbulence-Aware Decoder (TAD) to enhance temporal coherence while selectively mitigating turbulence-induced distortions.

Turbulence Mask Gating Given the latent feature z_t at time step t , we first apply temporal convolutions to extract inter-frame dependencies. To identify turbulence-corrupted regions, we construct a disturbance heatmap T_{map} based on bidirectional warping errors:

$$T_{\text{map}} = \|\text{Warp}(x_{t-1}, f_{t \rightarrow t-1}) - x_t\|_1 + \|\text{Warp}(x_{t+1}, f_{t \rightarrow t+1}) - x_t\|_1, \quad (9)$$

where $f_{t \rightarrow t \pm 1}$ denotes bidirectional optical flows estimated by the PhasorFlow module. The heatmap is converted to a gating mask $G \in [0, 1]^{H \times W}$ via $G = \sigma(\text{Conv}_{1 \times 1}(T_{\text{map}}))$, which modulates the temporal convolution output in a residual manner as:

$$f_t = \text{TMG}(z_t) = G \circ \text{Conv}_{1 \times 1}(\text{ResBlock}(z_t)). \quad (10)$$

This mechanism adaptively filters out turbulence-corrupted regions, ensuring that cross-frame modeling is restricted to structurally stable areas.

IR Structure-Aware Attention To further enhance the temporal alignment of critical structures, we introduce IR-SAA, which selectively enforces consistency in high-frequency regions (e.g., edges, contours) while avoiding redundant alignment in low-saliency regions.

From the output f_t of TMG, we construct a structure attention map $A_t \in [0, 1]^{H \times W}$ using the gradient magnitude as $A_t = \sigma(\text{Conv}_{1 \times 1}(\|\nabla f_t\|_1))$, and enhance the feature via residual attention $f_t^{\text{enh}} = f_t + \lambda(f_t \circ A_t)$, where λ is a fixed scaling coefficient, this allows the model to focus computational capacity on thermally relevant structures.

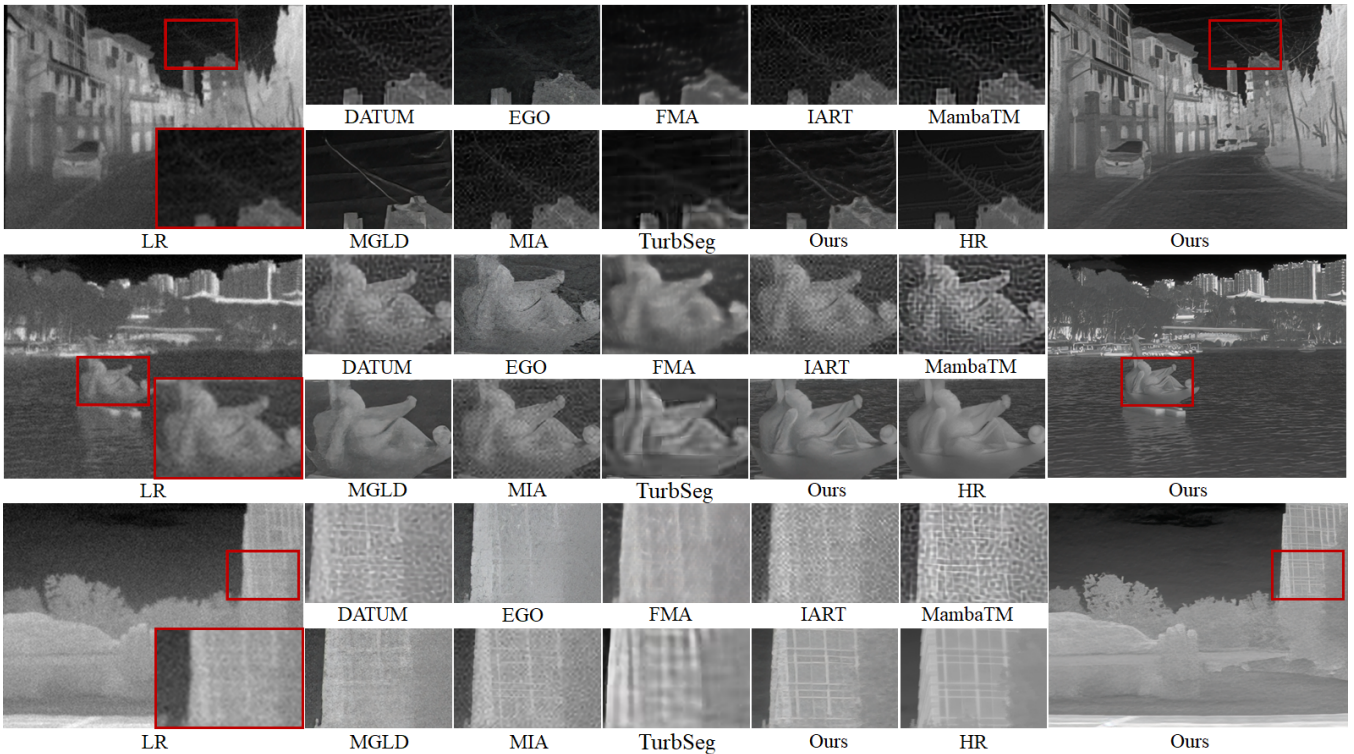


Figure 4: Qualitative results. The first row is from the static scenes of the M³FD dataset, while the second and third rows are from the FLIR-IVSR dataset. MambaTM, DATUM, and Turb-Seg are combined with BasicVSR to form a two-stage pipeline.

Optimization We fine-tune the TAD on top of a pre-trained VAE decoder for turbulent infrared VSR tasks. We first define the Thermal Reconstruction Loss to emphasize high-fidelity recovery in thermally active regions as $\mathcal{L}_{\text{thermal}} = \left\| (\hat{\mathbf{I}} - \mathbf{I}_{\text{gt}}) \circ M_{\text{phasor}} \right\|_1$, where M_{phasor} is Phasor Mask from thermal phasor analysis. To encourage sharper recovery of blurred thermal contours, we introduce the Thermal Edge Loss as $\mathcal{L}_{\text{edge}} = \left\| (\nabla \hat{\mathbf{I}} - \nabla \mathbf{I}_{\text{gt}}) \circ M_{\text{phasor}} \right\|_1$, where $\nabla(\cdot)$ denotes a Laplacian operator applied for edge extraction to penalize misalignment in thermal edge structures. Also, to preserve temporal consistency across the reconstructed sequence, we employ a Frame Difference Loss defined as $\mathcal{L}_{\text{difr}} = \sum_i \left\| (\hat{\mathbf{I}}_{i+1} - \hat{\mathbf{I}}_i) - (\mathbf{I}_{i+1}^{\text{gt}} - \mathbf{I}_i^{\text{gt}}) \right\|_1$.

The total loss function is then formulated by combining those loss functions. This joint loss not only enhances restoration in thermal-sensitive regions but also improves stability of the overall diffusion trajectory under turbulence.

Experiments

Experimental Settings

Implementation Details Our network is trained on an NVIDIA A800 GPU using the Adam optimizer, with hyper-parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We first fine-tune the U-Net backbone, initializing it with pretrained weights from Stable Diffusion v2.1 (Rombach et al. 2022). To ef-

fectively incorporate information from LR inputs, we introduce a lightweight time-aware encoder that extracts temporal features from LR images and encodes them as conditional inputs to guide the diffusion process. Subsequently, we train the proposed PhasorFlow module independently and integrate it with the fine-tuned U-Net to perform image sampling, which generates latent features for training the Turbulence-Aware Decoder.

Datasets and Evaluation Metrics To facilitate research in infrared video super-resolution under atmospheric turbulence, we construct FLIR-IVSR, an infrared VSR dataset comprising 640 paired LR-HR infrared video sequences captured using a FLIR T1050sc thermal camera at a resolution of 1024×768 . The dataset encompasses a wide range of motion patterns and scene categories, and is divided into two subsets based on camera motion. The camera-moving subset contains 135 sequences, featuring scenarios with platform-induced motion. The camera-static subset includes 510 sequences, further categorized into: (i) Dynamic scenes (495 sequences), characterized by object-level or environmental motion with a stationary camera; (ii) Static scenes (15 sequences) with minimal motion. FLIR-IVSR provides a comprehensive and challenging benchmark for assessing infrared VSR methods under severe low-resolution and turbulence-induced degradations. The process of building the FLIR-IVSR is detailed in the supplementary materials.

We train all models on the FLIR-IVSR training set, which

Datasets	Set5					Set10					Set20				
	Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	VMAF \uparrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	VMAF \uparrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow
M³FD															
MambaTM	25.6757	0.5741	0.4078	0.2319	28.0380	25.6237	0.5822	0.4084	0.2231	26.7815	25.6078	0.5779	0.4095	0.2299	26.3855
Turb-seg	22.1135	0.6857	0.2976	0.2402	5.7285	24.2823	0.7399	0.2582	0.2084	8.4070	23.7615	0.7361	0.2584	0.2152	6.1068
DATUM	<u>28.1310</u>	0.6749	0.3569	0.1880	45.8987	<u>28.3336</u>	0.7020	0.3420	0.1771	<u>45.6202</u>	<u>28.4232</u>	0.7026	0.3389	0.1807	<u>46.3185</u>
MGLDVSR	27.1603	<u>0.8003</u>	<u>0.2106</u>	0.1513	26.7114	27.9049	<u>0.7965</u>	<u>0.1919</u>	<u>0.1612</u>	25.5742	28.1681	<u>0.8137</u>	<u>0.1902</u>	0.1515	27.5137
FMA-NET	27.5482	0.7831	0.2376	0.2200	31.3568	27.0874	0.7784	0.2344	0.2105	29.4062	27.1545	0.7788	0.2324	0.2139	28.9050
MIA-VSR	27.7264	0.6153	0.3529	0.2461	44.9468	27.6816	0.6240	0.3576	0.2533	45.1764	27.7188	0.6221	0.3599	0.2534	44.9845
IART	27.7020	0.6020	0.3528	0.2542	<u>45.6605</u>	27.6319	0.6114	0.3576	0.2607	45.5397	27.6641	0.6089	0.3605	0.2608	45.3884
EGOVSR	26.6591	0.7230	0.2611	0.1975	27.1438	26.2876	0.7055	0.2865	0.1971	25.0401	26.3767	0.7102	0.2833	0.1992	24.9732
Ours	29.7819	0.8311	0.1724	<u>0.1576</u>	44.6731	30.6093	0.8352	0.1455	0.1479	48.6273	30.3834	0.8370	0.1555	<u>0.1530</u>	46.6925
FLIR-IVSR															
MambaTM	22.7972	0.3114	0.6511	0.3369	11.6541	23.3786	0.3256	0.6693	0.3654	12.8628	23.7571	0.3665	0.6267	0.3399	18.0775
Turb-seg	24.8976	<u>0.7509</u>	<u>0.2973</u>	0.2346	4.6408	23.0295	<u>0.7825</u>	<u>0.2770</u>	0.2559	4.3775	20.2981	0.6894	0.6375	0.3606	4.3461
DATUM	27.6349	0.5596	0.5063	0.2674	25.9121	27.9964	0.5688	0.5230	0.2981	24.4874	27.1081	0.5550	0.5156	0.2831	29.4297
MGLDVSR	29.2938	0.6336	0.3679	<u>0.2274</u>	28.1148	<u>30.4112</u>	0.7045	0.3519	<u>0.2476</u>	27.9592	27.5376	<u>0.7983</u>	<u>0.2072</u>	<u>0.1608</u>	25.5895
FMA-NET	<u>29.6184</u>	0.7457	0.3177	0.2618	28.6511	29.5584	0.7773	0.2843	0.2741	23.4312	28.0662	0.7261	0.3244	0.2710	26.2214
MIA-VSR	27.2881	0.4882	0.5169	0.3515	25.8345	27.5797	0.4920	0.5244	0.3752	24.1688	27.1045	0.4927	0.5171	0.3625	29.9711
IART	27.2596	0.4893	0.5008	0.3573	26.4507	27.5574	0.4940	0.5018	0.3815	24.7818	27.0212	0.4877	0.5024	0.3697	30.1541
EGOVSR	28.7845	0.6452	0.3835	0.2629	<u>36.7992</u>	29.5250	0.6645	0.4177	0.2938	<u>35.5069</u>	<u>28.4134</u>	0.6573	0.3873	0.2733	<u>32.1390</u>
Ours	33.3719	0.8683	0.1227	0.1183	46.6922	33.8680	0.8545	0.1555	0.1559	42.8454	32.4682	0.8415	0.1377	0.1464	44.8895

Table 1: Quantitative comparison on M³FD and FLIR-IVSR. The best is in **bold**, while the second is underlined. For M³FD, Set5/10/20 are randomly sampled subsets. For FLIR-IVSR, the three sets correspond to “camera-static (static scene)”, “camera-static (dynamic scene)”, and “camera-moving”, respectively.

consists of 505 turbulent infrared video sequence LR-HR pairs. Evaluation is conducted on two test sets: (1) the FLIR-IVSR test set comprising 135 turbulent infrared video pairs, and (2) a synthetic turbulence benchmark constructed from the static scenes of the public M³FD dataset by simulating turbulence-induced distortions.

To comprehensively assess both fidelity and perceptual quality, we report five widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Deep Image Structure and Texture Similarity (DISTS), and Video Multi-Method Assessment Fusion (VMAF). Detailed definitions of these metrics are provided in (Ma, Ma, and Li 2019).

Comparative Methods We perform a comprehensive comparison of our approach with five video super-resolution(VSR) methods, including MIA-VSR (Zhou et al. 2024b), FMA-Net (Youk, Oh, and Kim 2024), EGOVSR (Chi et al. 2024), IART (Xu et al. 2024), and MGLDVSR (Yang et al. 2024), as well as three turbulence removal methods, MambaTM (Zhang et al. 2025), DATUM (Zhang et al. 2024), and Turb-Seg (Saha et al. 2024). Notably, each turbulence removal method is combined with a unified VSR model, BasicVSR (Chan et al. 2022), forming two-stage pipelines that perform turbulence correction followed by resolution enhancement.

Qualitative Results

To visually demonstrate the effectiveness of our method, Figure 4 presents the restoration results of different approaches on the same frame of identical samples. The top sample comes from the turbulence-degraded M³FD dataset, while the bottom two are from our FLIR-IVSR dataset. As observed in the figure, the three two-stage ap-

Guide	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	VMAF \uparrow
PhasorFlow	33.6507	0.8535	0.1377	0.1482	45.3972
SpyNet	28.9387	0.7668	0.2386	0.1615	33.1466

Table 2: Quantitative ablation study on PhasorFlow.

proaches—MambaTM, DATUM, and Turb-Seg—that perform turbulence mitigation followed by super-resolution suffer from error accumulation during turbulence removal, and their subsequent super-resolution steps further amplify these artifacts. The four VSR methods—MIA-VSR, FMA-Net, EGOVSR, and IART—also fail to effectively address the noise, blurring, and spatial distortion caused by turbulence. While the diffusion-based VSR method MGLDVSR shows some capability in recovering blurred and noisy content, it still exhibits texture loss and restoration errors due to the lack of turbulence mitigation and infrared-specific guidance. In contrast, our approach successfully restores thermal details and spatial distortions while preserving high-resolution texture and maintaining the visual characteristics intrinsic to infrared imagery.

Quantitative Comparison

Table 1 compares the quantitative results on the FLIR-IVSR and turbulence-degraded M³FD datasets. For M³FD, subsets are randomly sampled for evaluation, while for FLIR-IVSR, test samples are selected from different scene categories to enable a comprehensive analysis. As demonstrated in the table, our method achieves the best performance among all compared approaches across different camera motions on the FLIR-IVSR dataset. For the M³FD dataset, we show clear advantages on the larger test sets (sizes 10 and 20), demonstrating the effectiveness of our method in mitigating complex degradation conditions for infrared VSR.

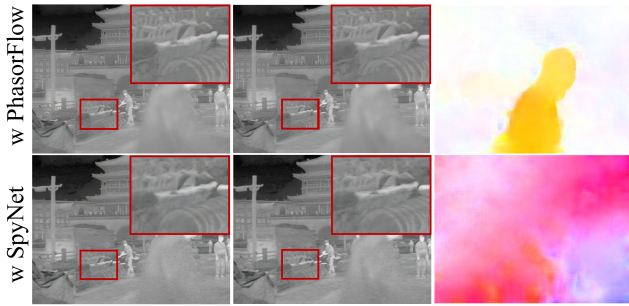


Figure 5: Qualitative ablation on the PhasorFlow.

IR-SAA	TMG	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	VMAF \uparrow
-	-	26.3283	0.6775	0.2862	0.1987	32.0941
\checkmark	-	27.3985	0.7169	0.1735	0.1735	36.2274
-	\checkmark	28.4125	0.7418	0.1564	0.1541	40.9598
\checkmark	\checkmark	32.2391	0.8229	0.1358	0.1431	43.4152

Table 3: Quantitative ablation on the TAD.



Figure 6: Qualitative ablation on the TAD.

Ablation Studies

Phasor-Guided Flow Estimator To validate the effectiveness of the proposed PhasorFlow, we replace it with the pre-trained optical flow network SpyNet (Ranjan and Black 2017). As shown in Table 2, PhasorFlow consistently outperforms SpyNet across all evaluation metrics, with notable improvements of approximately 4.7 dB in PSNR and 12 points in VMAF. These results demonstrate that PhasorFlow leads to significant enhancements in both structural fidelity and perceptual quality of the restored videos.

We further provide qualitative comparisons as illustrated in Figure 5. Compared with the results obtained using SpyNet, PhasorFlow better preserves object boundaries, produces clearer textures, and significantly suppresses background noise. These advantages are especially evident in thermally active regions, such as human silhouettes, where PhasorFlow provides more consistent and temporally stable flow fields. The corresponding optical flow maps further intuitively highlight its ability to preserve coherent motion boundaries in these regions, while SpyNet suffers from severe distortions and fragmented flow predictions.

Turbulence-Aware Decoder To evaluate the effectiveness of the Turbulence-Aware Decoder (TAD), we conduct an ablation study by removing its two key components: Turbulence Mask Gating (TMG) and IR Structure-Aware Attention (IR-SAA). As shown in Table 3, the absence of either module leads to noticeable performance drops. In particular, removing IR-SAA causes a significant decline in perceptual quality. In contrast, removing TMG primarily compromises

M_{occ}	M_{phasor}	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	VMAF \uparrow
-	-	26.3073	0.6558	0.3503	0.2370	34.0477
\checkmark	-	31.6965	0.7595	0.2866	0.2248	39.1762
-	\checkmark	28.9239	0.7149	0.2242	0.1817	30.4051
\checkmark	\checkmark	32.1595	0.8087	0.1573	0.1478	42.6042

Table 4: Quantitative ablation on the masked guidance.

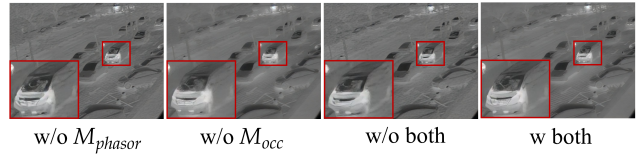


Figure 7: Qualitative ablation on the masked guidance.

alignment robustness and fidelity, as reflected by increased LPIPS and DISTS values. The removal of both modules leads to further degradation, underscoring the necessity of multi-level turbulence modeling for reliable restoration under severe distortions.

Figure 6 presents qualitative comparisons, demonstrating that the complete TAD yields richer texture details and more coherent background structures. These results highlight the complementary contributions of TMG and IR-SAA to structural modeling and consistency preservation.

Heat-Aware Guidance To validate the effectiveness of the Heat-Aware Guidance mechanism, we conduct a quantitative ablation study under four configurations: (1) without any heat-aware modulation mask, (2) using only the Phasor Mask M_{phasor} , (3) using only the Occlusion Mask M_{occ} , and (4) applying both to form the heat-aware modulation mask M_{joint} . As reported in Table 4, the joint application of both masks consistently achieves the best performance across all metrics, confirming their complementary roles in enhancing both perceptual quality and structural fidelity by localizing reliable temporal structures.

Figure 7 provides qualitative evidence. Without the heat-aware modulation mask, the restored images suffer from blurred contours and structure loss, particularly in fine-grained regions such as vehicle grilles.

Conclusion

We propose **HATIR**, a heat-aware diffusion framework that unifies alignment and restoration for turbulent infrared VSR. By introducing a phasor-guided flow estimator and a turbulence-aware decoder, HATIR integrates physically grounded priors into the denoising process, enabling robust structural recovery under severe turbulence. Experiments on the newly built FLIR-IVSR dataset validate the effectiveness of our approach.

Broader Impact HATIR enhances infrared VSR under turbulence, benefiting critical applications such as autonomous driving, surveillance, and thermal monitoring in low-visibility settings. The proposed FLIR-IVSR dataset encourages future research in infrared VSR.

Acknowledgments

This work was partially supported by the China Postdoctoral Science Foundation (2023M730741) and the National Natural Science Foundation of China (No.62302078).

References

- Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5972–5981.
- Chen, Z.; Long, F.; Qiu, Z.; Yao, T.; Zhou, W.; Luo, J.; and Mei, T. 2024. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9232–9241.
- Chi, Y.; Gu, J.; Zhang, J.; Yang, W.; and Tian, Y. 2024. EgoVSR: Towards High-Quality Egocentric Video Super-Resolution. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3897–3906.
- Huang, Y.; Wang, W.; and Wang, L. 2015. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Advances in neural information processing systems*, 28.
- Huang, Y.; Wang, W.; and Wang, L. 2017. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 1015–1028.
- Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020. Video super-resolution with recurrent structure-detail network. In *European conference on computer vision*, 645–660. Springer.
- Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3224–3232.
- Li, J.; Du, S.; Wu, C.; Leng, Y.; Song, R.; and Li, Y. 2022. Drcr net: Dense residual channel re-calibration network with non-local purification for spectral super resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1259–1268.
- Li, J.; Yu, H.; Chen, J.; Ding, X.; Wang, J.; Liu, J.; Zou, B.; and Ma, H. 2025a. A²RNet: Adversarial Attack Resilient Network for Robust Infrared and Visible Image Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4770–4778.
- Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 335–351. Springer.
- Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, 270–288. Springer.
- Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7534–7544.
- Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Gool, L. V. 2022. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35: 378–393.
- Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.
- Liu, J.; Li, X.; Wang, Z.; Jiang, Z.; Zhong, W.; Fan, W.; and Xu, B. 2024. PromptFusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*.
- Liu, J.; Zhang, B.; Mei, Q.; Li, X.; Zou, Y.; Jiang, Z.; Ma, L.; Liu, R.; and Fan, X. 2025a. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2226–2235.
- Liu, Y.; Zou, Y.; Li, X.; Zhu, X.; Han, K.; Jiang, Z.; Ma, L.; and Liu, J. 2025b. Toward a Training-Free Plug-and-Play Refinement Framework for Infrared and Visible Image Registration and Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1268–1277.
- Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45: 153–178.
- Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Saha, R. K.; Qin, D.; Li, N.; Ye, J.; and Jayasuriya, S. 2024. Turb-seg-res: a segment-then-restore pipeline for dynamic videos with atmospheric turbulence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25286–25296.
- Sajjadi, M. S.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6626–6634.

- Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3360–3369.
- Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.
- Wang, Y.; Jin, D.; Chen, J.; and Bai, X. 2023. Revelation of hidden 2D atmospheric turbulence strength fields from turbulence effects in infrared imaging. *Nature Computational Science*, 3(8): 687–699.
- Wang, Z.; Zhang, C.; Chen, Z.; Hu, W.; Lu, K.; Ge, L.; and Wang, Z. 2024. ACR-Net: Learning High-Accuracy Optical Flow via Adaptive-Aware Correlation Recurrent Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9064–9077.
- Wang, Z.; Zhang, J.; Guan, T.; Zhou, Y.; Li, X.; Dong, M.; and Liu, J. 2025a. Efficient Rectified Flow for Image Fusion. *Advances in Neural Information Processing Systems*.
- Wang, Z.; Zhang, J.; Song, H.; Ge, M.; Wang, J.; and Duan, H. 2025b. Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12637–12647.
- Xu, K.; Yu, Z.; Wang, X.; Mi, M. B.; and Yao, A. 2024. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2546–2555.
- Yang, X.; He, C.; Ma, J.; and Zhang, L. 2024. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European Conference on Computer Vision*, 224–242. Springer.
- Youk, G.; Oh, J.; and Kim, M. 2024. FMA-Net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 44–55.
- Zhang, X.; Chimitt, N.; Chi, Y.; Mao, Z.; and Chan, S. H. 2024. Spatio-temporal turbulence mitigation: a translational perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2889–2899.
- Zhang, X.; Chimitt, N.; Wang, X.; Yuan, Y.; and Chan, S. H. 2025. Learning Phase Distortion with Selective State Space Models for Video Turbulence Mitigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2127–2138.
- Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023. DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8082–8093.
- Zheng, M.; Sun, L.; Dong, J.; and Pan, J. 2025. Efficient Video Super-Resolution for Real-time Rendering with De-coupled G-buffer Guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11328–11337.
- Zhou, S.; Yang, P.; Wang, J.; Luo, Y.; and Loy, C. C. 2024a. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2535–2545.
- Zhou, X.; Zhang, L.; Zhao, X.; Wang, K.; Li, L.; and Gu, S. 2024b. Video super-resolution transformer with masked inter&intra-frame attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25399–25408.
- Zou, Y.; Chen, Z.; Zhang, Z.; Li, X.; Ma, L.; Liu, J.; Wang, P.; and Zhang, Y. 2026. Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution. *International Journal of Computer Vision*.
- Zou, Y.; Li, X.; Jiang, Z.; and Liu, J. 2024. Enhancing neural radiance fields with adaptive multi-exposure fusion: A bilevel optimization approach for novel view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7882–7890.