Learning Continuous Spatiotemporal Implicit Neural Fields for Unsupervised Video Denoising

Xiaowan Hu, *Member, IEEE*, Henan Liu, Ce Zheng, Xinyang Li, Mai Xu, *Senior Member, IEEE*

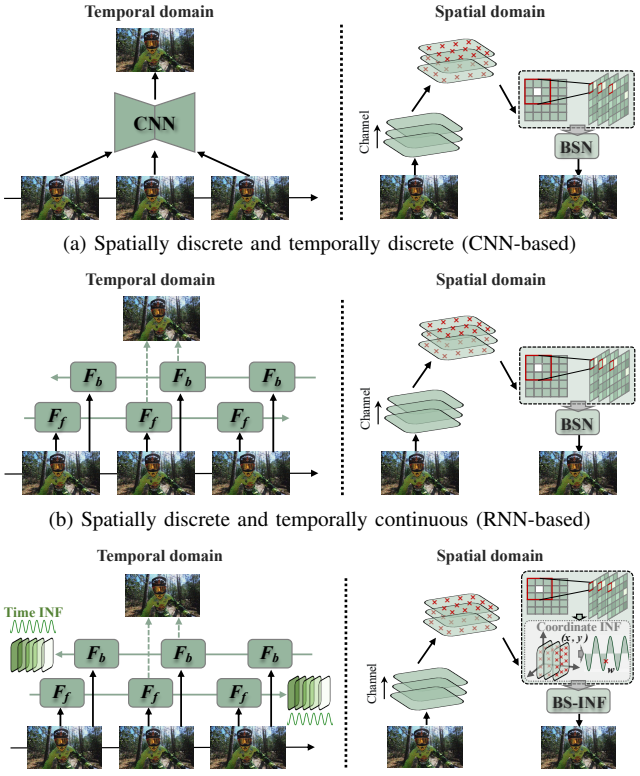
***Abstract*—Video denoising is fundamental to low-level vision and real-world imaging, yet existing self-supervised methods remain fragile under severe noise and complex motion. Most approaches still rely on spatially and temporally discrete grid-based representations: blind-spot networks enforce J-invariance by masking center pixels with a limited receptive field, while recurrent models build temporal dependencies on discretized frame sequences and noise-sensitive optical flow, leading to error accumulation and motion artifacts. We address this model bottleneck by reformulating self-supervised video denoising as learning a continuous spatiotemporal implicit field. Building on coordinate-based implicit neural representations, we propose a unified video denoising model with a spatiotemporal implicit neural field (SINF). In the spatial domain, a blind-spot implicit spatial field maps coordinates directly to pixel-level representations, enabling globally informed texture recovery beyond receptive-field limits. In the temporal domain, an implicit temporal embedding with periodic activations encodes motion continuously over time, while a time-aware spatial graph module refines cross-frame alignment. Together, SINF remodels discretized video signals into a continuous spatiotemporal intensity field, enabling more robust pixel-wise associations than coarse optical flow. Extensive experiments on synthetic and real noisy video benchmarks demonstrate that our SINF achieves state-of-the-art performance on synthetic and real noisy video benchmarks. Our code is publicly available at: https://github.com/Huxiaowan/SINF-VDN.**

***Index Terms*—Video denoising, self-supervised learning, implicit neural representation, spatiotemporal modeling.**

I. Introduction

VIDEO denoising is a key cornerstone of low-level vision and underpins numerous real-world applications [1]–[4]. However, obtaining well-aligned clean–noisy video pairs in dynamic scenes is challenging, which fundamentally limits the practical scalability of supervised learning approaches [5]–[8].

Self-supervised video denoising methods that learn solely from noisy inputs have attracted increasing research attention in recent years. Early studies predominantly borrowed ideas from self-supervised image denoising frameworks such as Noise2Noise [9] and Noise2Void [10], and evolved into two major technical paradigms. The first category, frame-registration based methods, constructs training pairs via noisy pairs or motion-compensated frame alignment (e.g., optical flow estimation) to enable self-supervision [11], [12]. The second category, blind-spot constrained spatiotemporal modeling methods, predicts the target frame by masking the center pixel and/or neighboring frames [5], [6]. These approaches have driven video denoising from single-frame to multi-frame and from spatial to spatiotemporal modeling, and now form the mainstream paradigm in unsupervised video denoising.



(a) Spatially discrete and temporally discrete (CNN-based)

(b) Spatially discrete and temporally continuous (RNN-based)

(c) Spatially and temporally continuous (INR-based)

Fig. 1: Three paradigms of self-supervised video denoising with different spatiotemporal continuity. (a) CNN-based methods operate on discretized frames with limited receptive fields; (b) RNN-based methods incorporate temporal continuity with spatial discretization; (c) INR-based methods treat video as a continuous and coherent function over both space and time.

Despite notable progress, existing self-supervised video denoising methods still exhibit two intrinsic bottlenecks, especially under severe noise and complex motion. First, the blind-spot network (BSN), while satisfying the J-invariance assumption for self-supervision, explicitly mask the center pixel, which restricts the effective receptive field and hinders the recovery of fine-grained texture details. Second, methods that rely on optical flow for frame registration are inherently sensitive to noise; flow estimation errors are easily amplified and temporally propagated, leading to motion artifacts and

Xiaowan Hu, Henan Liu, Ce Zheng, and Mai Xu are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: huxiaowan@buaa.edu.cn; lhn21373089@buaa.edu.cn; zhengce@buaa.edu.cn; MaiXu@buaa.edu.cn).

Xinyang Li is with the College of AI, Tsinghua University, Beijing, China (e-mail:xinyangli@tsinghua.edu.cn).

✉ Corresponding author: Mai Xu, Xinyang Li.

structural distortions. Prior work [1], [5], [13] has clearly demonstrated that the noise sensitivity of optical flow is a primary factor contributing to performance degradation.

To better contextualize these limitations, we categorize existing blind-spot video denoising methods by their treatment of spatial and temporal continuity. *(a) Spatially discrete and temporally discrete* architectures, as shown in Fig. 1a, rely on CNNs operating on fixed pixel grids and independent frame sequences; they can only aggregate local context and are inherently constrained by narrow receptive fields. *(b) Spatially discrete and temporally continuous* methods, depicted in Fig. 1b, extend temporal modeling with recurrent units such as RNNs, capturing longer-range temporal correlations. However, they still depend on discretized spatial grids and often require precise optical flow alignment, making them vulnerable to severe noise and fast motion. Recent recurrent self-supervised schemes [6], [12], [14] employ bidirectional temporal models to aggregate broader context but exhibit temporal discontinuities in highly dynamic scenes. Moreover, frequent down- and up-sampling operations used to decorrelate spatial noise introduce irreversible structural information loss. Overall, these prevailing discrete, grid-based architectures lack the capacity to provide truly continuous and coherent spatiotemporal modeling under realistic video conditions.

These findings motivate a shift in perspective: rather than further refining discrete grids, we seek a solution space that is continuous in both space and time. Recent advances in implicit neural representations (INR) [15]–[17] encode signals as continuous, differentiable functions of their coordinates, providing a principled alternative. Thus, we adopt an INR-based paradigm in Fig. 1c for *(c) Spatially and temporally continuous* video denoising. The INR-based framework treats video as a continuous spatiotemporal function by directly mapping space–time coordinates to pixel intensities or latent features. This coordinate-driven formulation naturally captures robust long-range alignment, reduces reliance on noise-sensitive optical flow, and is inherently resolution-agnostic. The differentiable nature of INR supports the formation of a coherent motion field, and periodic activation functions [15] provide strong spectral approximation capability for high-frequency detail recovery. Together, these properties make INR a compelling foundation for self-supervised video denoising.

Building on this insight, we propose spatiotemporal implicit neural fields (SINF) for unsupervised video denoising, introducing continuous spatiotemporal modeling into the self-supervised setting. In the spatial domain, we employs a blind-spot implicit neural field (BS-INF) that directly maps spatial coordinates to pixel representations while enforcing blind-spot constraints by excluding the center coordinate from queries, thereby overcoming the limitation of local receptive fields and enabling globally informed texture recovery. In the temporal domain, an implicit temporal embedding injects normalized timestamps into an implicit representation with periodic activations to learn a continuous mapping from time to feature space, implicitly encoding motion trajectories and high-frequency temporal variations without explicit optical flow. To tightly couple space and time, we further design a time-aware spatial graph module that constructs pixel-level graphs within local windows, aggregates similar features via spatial attention, and refines motion trajectories through cross-frame interaction with residual boundary compensation. Together, these components realize a unified continuous spatiotemporal reference frame, where spatial coordinates localize pixel positions and temporal coordinates encode motion states, substantially reducing reliance on accurate flow estimation. SINF uses coordinate-based implicit fields to model the noisy image formation process and thereby enabling robust self-supervised video denoising even under extremely challenging conditions. The main contributions are summarized as follows:

- We reformulate self-supervised video denoising from a discrete grid-based problem into learning a continuous spatiotemporal implicit field, yielding a unified coordinate-driven view of video in space and time.
- The blind-spot implicit spatial field directly maps spatial coordinates to pixel-level representations, eliminating convolutional blind-spot constraints and enabling globally informed texture recovery beyond receptive-field limits.
- An implicit temporal embedding with periodic activations encodes motion continuously over time, while a time-aware spatial graph module refines cross-frame alignment for coherent denoising without explicit optical flow.
- Extensive experiments on synthetic and real noisy video benchmarks demonstrate the state-of-the-art performance of our SINF, with improved detail preservation and fewer motion artifacts in highly noisy and dynamic scenarios.

II. Related Work

A. Supervised Video Denoising.

Early supervised video denoising techniques were largely inspired by image denoising methods, evolving from non-local similarity exploitation approaches such as Non-Local Means (NLM) [18], block-matching and 3D filtering (BM3D) [19], and sparse prior modeling [20], which extended spatial patch aggregation into the temporal domain but suffered from limited motion handling capabilities under complex dynamics. With the rise of deep learning, supervised video denoising shifted towards learning spatiotemporal priors directly from paired noisy–clean datasets, where many methods rely on explicit inter-frame alignment to ensure temporal coherence. Optical flow-based alignment, as in PWC-Net [21], warps adjacent frames before feature fusion, while deformable convolution-based alignment [22] enables spatially adaptive sampling to cope with local geometric variations. Representative examples include DVDnet [1] and FastDVDnet [13], which extend single-image denoisers into sliding-window architectures that jointly process multiple neighboring frames without explicit flow computation, and EDVR [23] and RViDeNet [24], which adopt pyramid-based, deformable alignment modules to handle large motion and parallax before multi-scale aggregation. Beyond local temporal windows, attention-based networks [25], [26] employ non-local or transformer architectures to capture long-range dependencies, while recurrent architectures like FloRNN [27], BasicVSR [28], and BasicVSR++ [29] leverage bidirectional or recursive propagation for progressive refinement with reduced memory cost. Relatedly, MSTMN [30] use

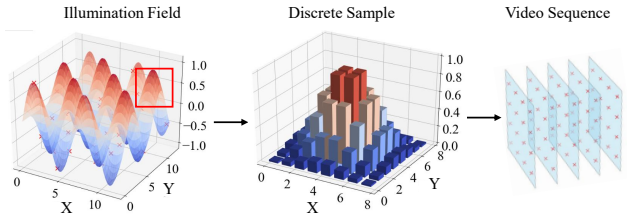
multi-scale spatiotemporal memory network for lightweight video denoising. Nevertheless, these approaches require paired clean–noisy video data and often do not generalize well to noisy real-world scenarios with unknown noise statistics.

B. Unsupervised Video Denoising

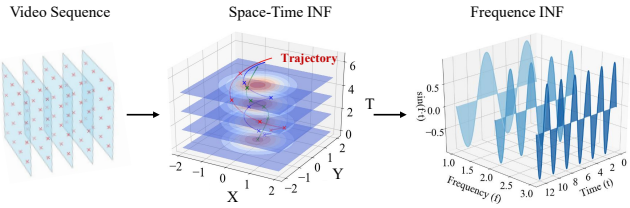
To remove the dependence on paired clean videos, a variety of unsupervised denoising frameworks have been proposed. A first line of work extends the Noise2Noise paradigm [9] to videos by exploiting noise independence across frames. Multi-Frame2Frame [11] uses motion-compensated neighbors to construct pseudo noisy–target pairs, while R2R [31] and VER2R [12] apply re-corruption schemes to synthesize multiple noise realizations of the same content, enabling learning from corrupted–corrupted pairs without clean supervision. A second line relies on blind-spot networks [32], which mask the receptive-field center to prevent identity mappings and thus allow self-supervised training from single noisy inputs. Complementary blind-spot designs have also been explored for real-noise self-supervised denoising [33], [34]. ViDeNN [35] predicts clean frames directly from noisy sequences, and UDVD [5] extends blind-spot modeling to videos by stacking neighboring frames for joint spatiotemporal denoising. Subsequent methods such as RDRF [6] and STBN [14] enrich recurrent propagation and spatial receptive fields to capture stronger temporal context, while TAP [6] adopts a teacher–student strategy that distills pre-trained image or video denoisers as soft supervision on the target domain. Beyond natural videos, self-supervised denoising has also been applied to fluorescence microscopy and volumetric imaging under extremely low signal-to-noise ratios. DeepCAD [3], [36] and DeepSeMi [37] leverages repeated or temporally adjacent measurements in 3D calcium imaging, and SRDTrans [38] exploits spatial redundancy with transformer-based modeling to recover fine neural structures. Yet these methods still rely on short temporal windows, grid-based representations, and explicit frame alignment, making them sensitive to flow errors and illumination changes and limiting fine-detail recovery.

C. Implicit Neural Representations

Implicit neural fields (INF) [15], [17] have recently emerged as a powerful framework, which model continuous signals for coordinate-conditioned representations, in contrast to grid-based CNNs that operate on discrete pixels or voxels. Early works such as SIREN [15] showed that simple MLPs, driven only by spatial coordinates, can faithfully fit complex geometry and high-frequency patterns through suitable activation schemes. Building on this idea, LIIF [39] and MetaSR [40] couple feature encoders with local implicit heads to realize arbitrary-scale image super-resolution, decoupling output resolution from the training grid. INR-based image restoration models further exploit the implicit smoothness and spectral bias of coordinate MLPs for unsupervised or instance-specific recovery, particularly for blind image restoration [41], [42]. In dynamic scene modeling, NeRF [16] and time-aware variants Boosting-NeRV [43] jointly parameterize radiance as a function of 3D position, yielding temporally consistent novel-view synthesis. For video restoration, VideoINR [44] represents a video as a space–time neural field for joint spatial and temporal



(a) Continuous field to discrete video.



(b) Discrete video to spatiotemporal and spectral INFs.

Fig. 2: Formulation illustration: (a) a video as discrete samples from a continuous illumination field; (b) the continuous spatiotemporal and spectral implicit field from discrete videos.

super-resolution, while VRINR [45] extends this idea to a general video restoration backbone that avoids explicit optical-flow estimation. These studies collectively indicate that INRs naturally impose spatiotemporal continuity and cross-frame consistency without relying on discrete grids or fragile motion estimation. However, INR-based designs specifically tailored for self-supervised video denoising remain largely unexplored.

III. Preliminary

In this section, we provide the theoretical motivation for our design from three complementary perspectives. First, we view video as a discrete sampling of an underlying continuous illumination field and show how fixed space–time grids constrain representational capacity. Second, we recall the approximation properties of implicit neural fields, which offer expressive and resolution-agnostic modeling of spatiotemporal signals. Third, we exploit their differentiability to access spatial and temporal derivatives, enabling more robust modeling of motion dynamics and spatiotemporal consistency. These three aspects are formalized as core principles that guide the SINF framework.

A. Discrete Video Modeling

Proposition 1. *The video is a discrete sampling of the continuous illumination field $L(x, y, t)$ on a space-time grid. This representation has limited capacity at fixed resolution.*

Explanation. Traditional video modeling methods represent videos as a sequence of discrete image frames, where each frame is a 2D grid of pixel intensities. Let the ideal video signal be a continuous spatiotemporal function:

$$L(x, y, t) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3, \tag{1}$$

where $(x, y) \in \mathbb{R}^2$ are continuous spatial coordinates, $t \in \mathbb{R}$ is continuous time, and the output in $\mathbb{R}^3$ denotes RGB intensity. This formulation assumes that the visual scene evolves smoothly over both space and time, capturing geometric and photometric variations, as illustrated in Fig. 2a.

However, in practical acquisition systems, $L(x, y, t)$ is sampled over a discrete space-time lattice $V(i, j, k)$:

$$V(i, j, k) = L(x_j, y_k, t_i), \tag{2}$$

where $x_j = j \cdot \Delta x$, $y_k = k \cdot \Delta y$, $t_i = i \cdot \Delta t$ with Δx, Δy, and Δt being the spatial and temporal sampling intervals.

The fixed-resolution discrete grid thus imposes a strict upper bound on representational capacity. Formally, the sampled video can be expressed as a projection:

$$\Pi_{\mathcal{F}_N} L = \sum_{p=1}^{P} c_p \, \phi_p(x, y, t), \tag{3}$$

where $\Pi_{\mathcal{F}_N}$ denotes the orthogonal projection from $\mathcal{F}$ onto $\mathcal{F}_N$, and c_p is defined by the sampled values. The different frequency content is approximated by the limited basis $\{\phi_p\}$. Since $\dim(\mathcal{F}_N) = P$ is finite, any component of L lying outside $\mathcal{F}_N$ cannot be represented without error, including fine-grained spatial detail or high-order temporal variation.

This finite-dimensional constraint implies that, for a fixed spatial resolution (N_x, N_y) and temporal resolution N_t, the representational capacity of a grid-based video model is strictly bounded. Even modern deep architectures cannot inherently recover the continuous trajectories of moving objects or high-order temporal variations. High-frequency content in either space or time is only approximated by the limited basis.

Modeling the illumination field $L(x, y, t)$ as a continuous function can break free from the constraints of fixed grids. The shift towards continuous-domain representations is essential.

B. Implicit Neural Field

Proposition 2. *An implicit neural field $f_\theta(x, y, t)$ can approximate a continuous video function with arbitrary precision, offering superior frequency modeling capability.*

Explanation. Let the underlying video signal be a continuous spatiotemporal function:

$$V : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3, \quad L(x, y, t) \mapsto V(x, y, t), \tag{4}$$

where (x, y) denote continuous spatial coordinates and t denotes continuous time. An implicit neural field is defined as a neural network parameterized by θ:

$$f_\theta : \mathbb{R}^3(x, y, t) \to \mathbb{R}^3(r, g, b), \tag{5}$$

which directly maps normalized coordinates (x, y, t) to their corresponding RGB intensity values (r, g, b). As illustrated in Fig. 2b, discrete video samples supervise the spatiotemporal implicit field, which we further analyze in the spectral domain.

Unlike explicit grid-based methods that store discrete samples $V(i, j, k)$ in memory, the implicit formulation treats the video as a continuous function over space and time. This continuous view enables evaluation at arbitrary resolutions and time steps, which is naturally suited to restoration tasks. Its advantages can be summarized as follows.

1) Universal approximation. From a theoretical perspective, the universal approximation theorem ensures that, for any continuous video function V and any $\varepsilon > 0$, there exists a parameter set θ such that:

$$\sup_{(x,y,t)\in\Omega} \|f_\theta(x, y, t) - V(x, y, t)\|_2 < \varepsilon, \tag{6}$$

provided f_θ has sufficient depth, width, and non-linear activation. Here, $\Omega \subset \mathbb{R}^3$ denotes the bounded spatiotemporal domain of interest. Thus, an implicit neural field can approximate the underlying video signal with arbitrarily small error.

2) High-frequency expressivity. Conventional neural networks with ReLU or other low-frequency-biased activations tend to underfit rapidly varying signals due to spectral bias. By contrast, applying Fourier feature embeddings:

$$\gamma : \mathbb{R}^3 \to \mathbb{R}^{2m}, \quad \gamma(\mathbf{p}) = \big[\sin(2\pi \mathbf{B}\mathbf{p}),\ \cos(2\pi \mathbf{B}\mathbf{p})\big], \tag{7}$$

where $\mathbf{p} = (x, y, t)$ and $\mathbf{B} \in \mathbb{R}^{m\times 3}$ is a frequency scaling matrix, or adopting sinusoidal activations as in SIREN, allows the implicit neural field f_θ to represent signals with high spatial frequencies and fine temporal variations. This capability is crucial for reconstructing edges, textures, and fine-grained motion that are often lost in grid-based approximations.

3) Parameter efficiency. Implicit neural fields provide a compact parameterization. An explicit grid-based representation of a video of size $H \times W$ with T frames requires storing $\mathcal{O}(HWT)$ scalar values (e.g., RGB intensities for each spatiotemporal sample). In contrast, an implicit neural field $f_\theta : \mathbb{R}^3 \to \mathbb{R}^3$ is parameterized by a fixed-size vector θ of length $P(\theta)$, and the storage ratio between explicit and implicit representations can be expressed as:

$$\eta = \frac{P(\theta)}{HWT} \ll 1. \tag{8}$$

Moreover, since f_θ defines a continuous mapping over coordinates, its representation is independent of the discretization level, enabling evaluation at arbitrary resolutions without increasing the parameter count. This compactness supports generalization to unseen coordinates without retraining.

In general, an INF model embeds video as a continuous coordinate-to-signal map $f_\theta(x, y, t)$, replacing fixed-grid storage with a parametric spatial function. Leveraging universal approximation and frequency-rich embeddings, it provides a high-fidelity, resolution-independent representation and a principled alternative to discrete grid-based video models.

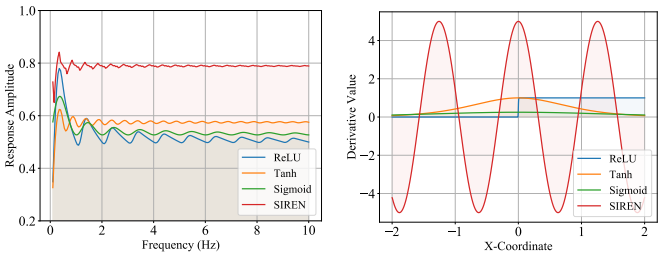
C. Spatiotemporal Continuous Learning

Proposition 3. *In video modeling, an everywhere differentiable function $\nabla_{x,y,t} f_\theta(x, y, t)$ enables robust and continuous learning of higher-order spatiotemporal dynamics.*

Explanation. Temporal coherence is critical in video tasks, yet optical-flow based methods often fail under noise, occlusions, or non-rigid motion. Implicit differential modeling offers a more robust alternative by letting the model infer motion directly from the derivatives of a continuous field. Since the INF model $f_\theta(x, y, t)$ is a continuous and smooth function, its temporal and spatial derivatives can be analytically computed and encode rich dynamic information:

$$\frac{\partial f_\theta}{\partial t}(x, y, t) \to \text{local velocity at pixel } (x, y), \tag{9}$$

$$\frac{\partial^2 f_\theta}{\partial t^2}(x, y, t) \to \text{acceleration or motion change rate.} \tag{10}$$



(a) Frequency response (b) Derivative smoothness

Fig. 3: Comparison of four activation functions: (a) Response amplitude across temporal frequencies; (b) Second derivative smoothness, showing the continuity advantage of SIREN.

Similarly, spatial gradients capture texture edges and sharpness, which are useful for structure preservation:

$$\nabla_{x,y} f_\theta(x, y, t) = \left(\frac{\partial f_\theta}{\partial x}, \frac{\partial f_\theta}{\partial y} \right). \tag{11}$$

The temporal evolution of a pixel value can then be formulated through Taylor expansion or integration of its derivatives. Under the assumption of smooth motion, the intensity at time t_2 can be approximated by:

$$f_\theta(x, y, t_2) \approx f_\theta(x, y, t_1) + \int_{t_1}^{t_2} \frac{\partial f_\theta}{\partial t}(x, y, \tau)\, d\tau, \tag{12}$$

which enables the reconstruction of pixel trajectories over time, promoting natural motion continuity and mitigating frame-to-frame inconsistency.

In practice, accurately capturing such trajectories requires activation functions with sufficient temporal bandwidth. In Fig. 3a, sinusoidal representation networks (SIREN) [15] exhibit a markedly stronger high-frequency response than ReLU, Tanh, and Sigmoid, making it well suited for modeling fast temporal variations. SIREN achieves this by using periodic activations, where each hidden layer is defined as:

$$\mathbf{h}_i = \sin\left(\omega_0 \mathbf{W}_i \mathbf{h}_{i-1} + \mathbf{b}_i\right), \tag{13}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters and ω_0 controls the frequency range. SIREN is infinitely differentiable, ensuring that f_θ and its temporal derivatives $\frac{\partial^n f_\theta}{\partial t^n}$ exist for all $n \in \mathbb{N}$. As sinusoids form a Fourier basis, the activation can represent both smooth low-frequency motion and sharp high-frequency variations. The temporal component of ∇f_θ is written as:

$$\nabla f_\theta(t) \approx \sum_{k=-K}^{K} a_k e^{j\omega_k t}, \quad \omega_k \in \mathbb{R}, \tag{14}$$

where the coefficients a_k are implicitly learned, and the set of frequencies $\{\omega_k\}$ is determined by ω_0 and the learned weights.

Beyond frequency coverage, as shown in Fig. 3b, SIREN exhibits smoother and more stable second derivatives, providing well-behaved higher-order dynamics. Such everywhere-differentiable implicit representations support continuous spatiotemporal modeling, where high-order motion cues enforce temporal coherence and preserve fine textures under noise.

IV. Method

A. Overview

In this chapter, the unsupervised video denoising pipeline was first formalized in IV-A1, and enabled by the blind-spot J-invariant paradigm in IV-A2. Our overall framework integrates bidirectional encoding, spatio-temporal implicit modeling, and graph alignment, optimizing a unified model to produce temporally consistent clean outputs in IV-A3 and IV-A4.

1) Problem Definition: We consider the task of unsupervised video denoising, where the input is a noisy sequence:

$$\mathcal{V} = \{I_t\}_{t=1}^{T}, \qquad I_t \in \mathbb{R}^{H \times W \times 3},$$

with T frames of spatial resolution $H \times W$. Each noisy frame I_t is assumed to be a degraded observation of the unknown clean frame C_t, corrupted by varying noise ϵ_t, formulated as

$$I_t(\mathbf{s}) = C_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad \mathbf{s} = (x, y) \in \Omega, \tag{15}$$

where $\epsilon_t(\mathbf{s})$ may follow a non-i.i.d. and heteroscedastic distribution, and Ω denotes the spatial domain of pixels.

The objective is to recover the clean video $\{C_t\}_{t=1}^{T}$ solely from the noisy observations $\{I_t\}$. Unlike grid-based methods, we model the video as a differentiable spatiotemporal field over $(\mathbf{s}, \tau) \in \mathbb{R}^2 \times \mathbb{R}$ and learn a parametric function

$$F_\Theta : (\mathbf{s}, \tau) \mapsto \widehat{\mathbf{c}} \in \mathbb{R}^3, \tag{16}$$

where $\widehat{\mathbf{c}}$ is the predicted RGB value at arbitrary coordinates. At integer frame indices $t \in \{1, \ldots, T\}$, we evaluate the field at $\tau = t$ to recover the discrete clean frame estimates $\widehat{C}_t$; at fractional τ, it enables temporally continuous interpolation.

From this perspective, the problem is equivalent to reconstructing an underlying clean signal manifold in $\mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3$ space, using only sparse and noisy samples $\{I_t\}$ observed on integer lattice points. The continuous formulation naturally bypasses the limited receptive field and optical flow reliance.

2) Self-Supervised Paradigm: Without paired labels, a core challenge in video denoising lies in preventing the network from degenerating into identity mapping, i.e., directly reproducing the noisy input without learning the underlying signal.

Blind-spot based self-supervision has emerged as a widely adopted paradigm. The central idea is to enforce *$\mathcal{J}$-invariance*: each prediction is inferred from spatiotemporal context without using its own noisy observation. We implement this by masking the target in the input and computing the loss only on masked sites, preventing leakage of the held-out measurement.

Formally, given a noisy frame I_t, we define a masking operator $\mathcal{M}$ that removes a subset of pixels from the receptive field of the network. Let $\mathbf{x} \in \Omega$ denote a spatial location, then the masked input is expressed as

$$I_t^{(\mathcal{M})}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \mathcal{M}, \\ I_t(\mathbf{x}), & \mathbf{x} \notin \mathcal{M}, \end{cases} \tag{17}$$

where $\mathcal{M}$ is sampled stochastically per iteration, ensuring that each pixel is eventually predicted without self-reference. For any masked target $\mathbf{x} \in \mathcal{M}$, the estimator can be written as:

$$\widehat{C}_t(\mathbf{x}) = f\big(I_t^{(\mathcal{M})}, \{I_{t-k}, \ldots, I_{t+k}\}\big)\Big|_{\mathbf{x}}, \qquad \frac{\partial \widehat{C}_t(\mathbf{x})}{\partial I_t(\mathbf{x})} = 0, \tag{18}$$

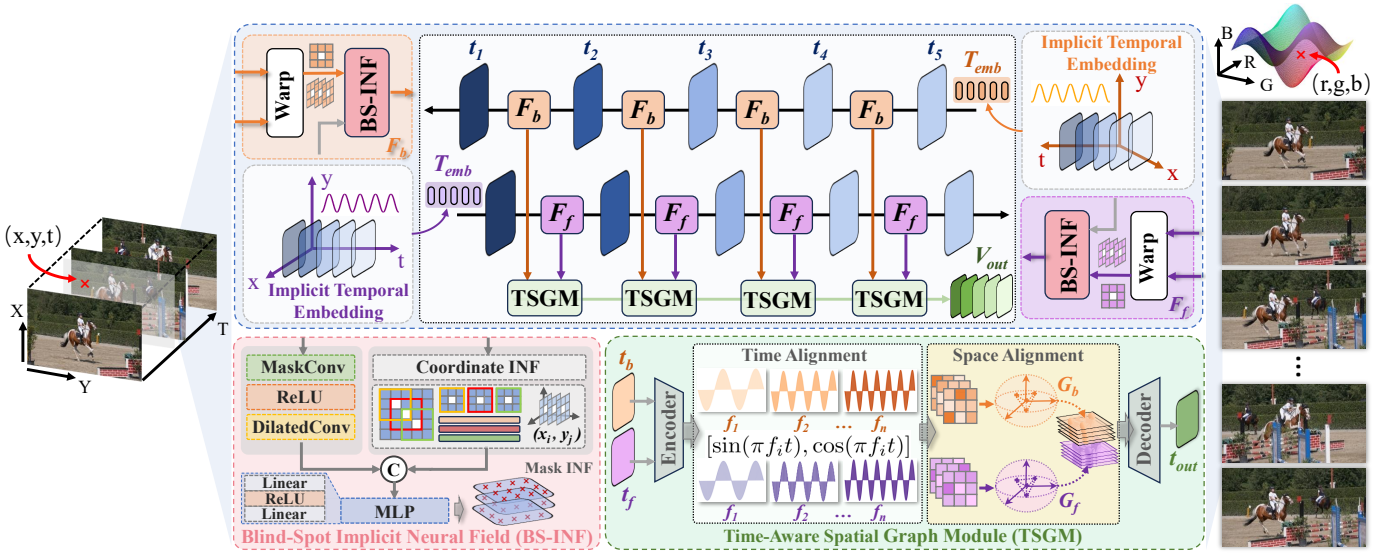


Fig. 4: Overall architecture of the proposed SINF model. Given a sliding window of noisy frames, a blind-spot INF expands the spatial receptive field via coordinate-based masking, an implicit temporal embedding provides continuous time encoding, and a time-aware spatial graph module refines cross-frame alignment, yielding denoised frames with spatiotemporal coherence.

meaning that the prediction at $\mathbf{x}$ is invariant to the held-out observation $I_t(\mathbf{x})$. Since the mask is resampled across iterations, taking expectation over random masks ensures that every pixel is optimized under the same blind-spot constraint.

Given the masked input $I_t^{(\mathcal{M})}$ together with adjacent frames $\{I_{t-k}, \ldots, I_{t+k}\}$, the network learns to reconstruct the clean estimate $\widehat{C}_t$ at the corresponding positions. Since the target pixel $I_t(\mathbf{x})$ is excluded from the receptive field, the model cannot trivially replicate the noisy observation, but is instead encouraged to exploit spatio-temporal correlations for restoration, enabling fully unsupervised learning.

3) Overall Network: As shown in Fig. 4, our framework builds on the blind-spot paradigm and integrates spatio-temporal information in a continuous and consistent manner. The goal is to learn an implicit function that maps discrete spatio-temporal coordinates (x, y, t) into RGB values, thereby reconstructing a clean sequence from noisy inputs. The blind-spot mask is applied at the network input, so all subsequent modules operate only on masked-context features and the center measurement is never accessible to the prediction.

Specifically, given a sliding window of consecutive frames:

$$\mathcal{I}_t = \{I_{t-K}, \ldots, I_t, \ldots, I_{t+K}\}, \quad I_t \in \mathbb{R}^{H\times W\times 3},$$

the network progressively refines the noisy center frame I_t into its clean counterpart $\widehat{C}_t$ while ensuring temporal coherence across the video. We perform bidirectional encoding with two complementary encoders, a forward encoder F_f and a backward encoder F_b, which take each masked frame $I_\tau^{(\mathcal{M})}$ and propagate information in opposite temporal directions. The self-supervised loss is evaluated on masked target, ensuring that each predicted pixel is trained without self-reference.

To enrich temporal awareness, a global implicit temporal embedding (ITE) maps the normalized timestamp τ/T into a continuous feature space via a sinusoidal activated MLP. The resulting time encoding is injected into both encoders, such that the encoded features are aware of both spatial content and temporal position. At each timestep, the forward and backward states are concatenated into a unified temporal representation:

$$\mathbf{H}_t = F_f(I_t^{(\mathcal{M})}, \mathrm{ITE}(t)) \oplus F_b(I_t^{(\mathcal{M})}, \mathrm{ITE}(t)). \tag{19}$$

Inside both F_f and F_b, the Blind-Spot Implicit Neural Field (BS-INF) is proposed as the core representation mechanism. The encoder feature maps are projected into coordinate-aware embeddings through a modulation function ψ:

$$\mathbf{z}_\tau(x, y) = \psi(\mathbf{F}_\tau(x, y)), \quad \mathbf{z}_\tau \in \mathbb{R}^{d_z}, \tag{20}$$

which are then combined with spatio-temporal positional encodings $\gamma(x, y, \tau)$ and decoded by the implicit function F_Θ to produce local blind-spot constrained predictions:

$$\widehat{\mathbf{c}}_\tau(x, y) = F_\Theta(\gamma(x, y, \tau), \mathbf{z}_\tau(x, y)). \tag{21}$$

Since $\mathbf{z}_\tau$ is extracted from masked inputs, the coordinate-based decoding does not introduce any access to the held-out center measurement. Embedding BS-INF into bidirectional encoders enables the network to enforce a global blind-spot receptive field, while building a continuous implicit coordinate-based representation that captures global spatio-temporal context.

Then, we designed the time-aware spatial graph module (TSGM) to achieve bidirectional spatio-temporal alignment and cross-frame fusion. For each spatial location (x, y) in the current frame, TSGM constructs a graph over its neighborhood $\mathcal{N}(x, y)$ and adjacent frames $\{t-K, \ldots, t+K\}$. Temporal alignment is guided by the continuous embeddings provided by ITE, while spatial alignment is achieved through graph-based attention that adaptively aggregates semantically consistent pixels. TSGM aggregates within a symmetric temporal window $\tau \in [t-K, t+K]$ of length $2K+1$, and applies boundary renormalization when the window is truncated. Nodes are window tokens in a local $w_s \times w_s$ region and edges are instantiated by localized attention between time coordinate,

with time weights computed from the normalized timestamp. The fused representation is formalized as:

$$\mathbf{z}_t^*(x,y) = \mathrm{TSGM}\big(\{\widehat{\mathbf{c}}_\tau(x',y')\}_{\tau\in[t-K,t+K],(x',y')\in\mathcal{N}(x,y)}\big). \tag{22}$$

The TSGM aims to aggregate contextual predictions within a bounded window and thus preserves the blind-spot constraint while enlarging the usable spatiotemporal context.

Finally, the fused embeddings $\mathbf{z}_t^*(x,y)$ are fed back into the implicit field F_Θ to reconstruct the clean output:

$$\widehat{C}_t(x,y) = F_\Theta(\gamma(x,y,t), \mathbf{z}_t^*(x,y)). \tag{23}$$

Collecting predictions over all spatial coordinates $(x,y) \in \Omega$ forms the clean frame $\widehat{C}_t$. And we repeated this process for each t reconstructs the full denoised video $\widehat{\mathcal{V}} = \{\widehat{C}_t\}_{t=1}^T$.

The entire pipeline can be summarized as

$$\mathcal{I}_t \xrightarrow{(F_f,F_b,\text{BS-INF})+\text{ITE}} \mathbf{H}_t \xrightarrow{\text{TSGM}} \mathbf{z}_t^* \xrightarrow{F_\Theta} \widehat{C}_t,$$

where F_f and F_b provide globally enriched blind-spot features, ITE offers continuous temporal awareness, TSGM achieves spatio-temporal alignment, and the implicit field reconstructs the final clean outputs. By integrating these components into a unified pipeline, the network progressively refines noisy observations into spatio-temporally consistent clean predictions.

4) Loss Function: The self-supervised loss is constructed by comparing network predictions with observed noisy pixels at masked positions. Let $\widehat{C}_t$ denote the reconstructed clean estimate of the center frame I_t, and let $\mathcal{M}$ denote the blind-spot mask defined in Eq. (17). The prediction of a pixel at location $\mathbf{x} \in \mathcal{M}$ is given by $\widehat{C}_t(\mathbf{x})$, while the corresponding noisy observation is $I_t(\mathbf{x})$. The blind-spot reconstruction loss for frame t is formulated as the masked mean squared error:

$$\mathcal{L}_t = \frac{1}{|\mathcal{M}|}\sum_{\mathbf{x}\in\mathcal{M}} \big\|\widehat{C}_t(\mathbf{x}) - I_t(\mathbf{x})\big\|_2^2. \tag{24}$$

Since supervision is imposed on multi-scale masked sites, the blind-spot objective does not assume i.i.d. Gaussian noise and remains compatible with non-i.i.d. heteroscedastic corruption.

Extending to the full sequence $\mathcal{V} = \{I_t\}_{t=1}^T$, the total training loss is obtained by averaging over all frames and their corresponding randomly sampled masks:

$$\mathcal{L}_{\text{total}} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{M}}\big[\mathcal{L}_t\big], \tag{25}$$

where the expectation $\mathbb{E}_{\mathcal{M}}[\cdot]$ is taken over randomly sampled mask patterns during training. It ensures that all pixels are eventually optimized, while strictly enforcing the J-invariance constraint inherent to self-supervised video denoising. No noise-type parameters are required in either training or inference, as optimization is driven solely by the blind-spot self-supervision on the observed sequence, enabling direct cross-domain transfer under native noise statistics.

B. Blind-Spot Implicit Neural Field

Blind-spot denoising masks the center pixel to avoid identity mapping but remains constrained by local receptive fields on discrete grids, struggles to capture motion, long-range dependencies, and structured noise. The BS-INF module couples blind-spot masking with spatiotemporal coordinate-aware implicit modeling, expanding the effective receptive field and recovering detailed context beyond convolutional limits.

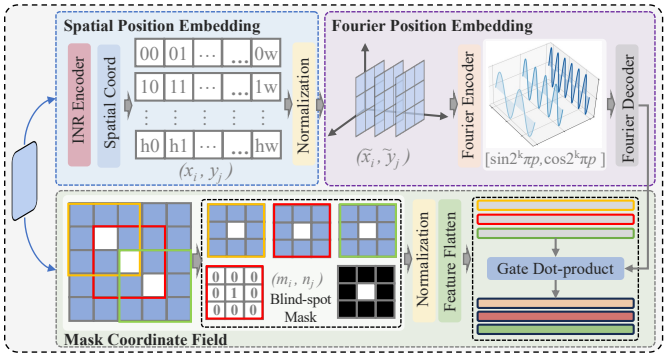


Fig. 5: Internal details of the coordinate INF module. The mask coordinate field is modulated by spatial and Fourier position embedding, producing globally consistent mask encoding.

As shown in Fig. 4, our overall architecture incorporates the BS-INF module, which integrates masked convolutions with coordinate-driven implicit neural representations to adaptively expand the receptive field over the entire image.

Given the flow-aligned feature map $\mathbf{F}_\tau \in \mathbb{R}^{H\times W\times d}$ corresponding to frame I_τ, we first apply masked convolutions that remove the center pixel from the receptive field, followed by dilated convolutions to enlarge neighborhood coverage:

$$\mathbf{U}_\tau = \mathrm{DilatedConv}\big(\mathrm{MaskConv}(\mathbf{F}_\tau)\big), \tag{26}$$

which aggregates detailed local structures around the masked location without directly accessing the noisy center pixel.

In parallel, as shown in Fig. 5, the coordinate-INF branch treats inputs as continuous signals by constructing a masked coordinate field and modulating it with spatial and Fourier positional embeddings. Each pixel index (x,y) in frame τ is first normalized into $[-1,1]$ to eliminate scale dependence:

$$\tilde{x} = \frac{2x}{W} - 1, \quad \tilde{y} = \frac{2y}{H} - 1, \quad \tilde{t} = \frac{2\tau}{T} - 1, \tag{27}$$

where (W,H,T) denotes the spatial width, height, and the total number of frames. Then we embed these normalized coordinates into a high-dimensional Fourier feature space:

$$\gamma(\mathbf{p}) = \big[\sin(2^b\pi\mathbf{p}), \cos(2^b\pi\mathbf{p})\big]_{b=0}^{B}, \quad \mathbf{p} = (\tilde{x},\tilde{y},\tilde{t}), \tag{28}$$

which expands $(\tilde{x},\tilde{y},\tilde{t})$ into multi-band sinusoidal basis functions, enriching fine-scale spatial and temporal structures.

To maintain blind-spot masking in the coordinate domain, we further incorporate a binary mask that indicates whether a location corresponds to a visible neighbor or a hidden center pixel. The masked positional encoding is defined as:

$$\tilde{\gamma}(x,y,\tau) = \mathbf{m}(x,y)\cdot\gamma(\tilde{x},\tilde{y},\tilde{t}), \tag{29}$$

ensuring that the implicit field is driven only by non-masked coordinates, while still allowing the network to infer missing center values through continuous spatiotemporal mapping.

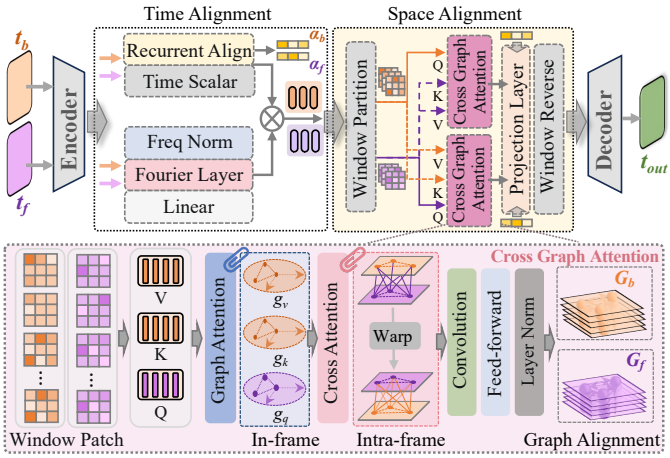


Fig. 6: Details of the time-aware spatial graph module. Time alignment encodes temporal coordinates, while space alignment aggregates local interaction via the cross graph attention.

The implicit neural field is parameterized by an MLP f_θ, which maps the encoded coordinates and local features into a continuous latent representation:

$$\mathbf{v}_\tau(x,y) = f_\theta\big(\tilde{\gamma}(x,y,\tau), \mathbf{F}_\tau(x,y)\big), \quad \mathbf{v}_\tau(x,y) \in \mathbb{R}^{d_v}. \tag{30}$$

This implicit representation effectively endows each blind-spot with a theoretically unbounded receptive field, enabling reconstruction of the masked center from global spatiotemporal dependencies rather than local neighborhoods alone.

The local masked–dilated features and global coordinate-INF features are then fused through a lightweight multilayer perceptron, consisting of a linear layer, a ReLU activation, and a second linear projection:

$$\mathbf{z}_\tau(x,y) = \mathrm{MLP}\big([\mathbf{U}_\tau(x,y), \mathbf{v}_\tau(x,y)]\big). \tag{31}$$

The fused blind-spot features $\mathbf{z}_\tau$ are propagated through the forward encoder F_f and backward encoder F_b for bidirectional prediction $\widehat{\mathbf{c}}_\tau(x,y)$ in Eq. (21), effectively balancing local masked aggregation and global implicit modeling.

C. Implicit Temporal Embedding

The implicit temporal embedding module provides explicit temporal position awareness through continuous coordinate representation. It preserves frame order, enhances long-range consistency, and offers a smooth temporal prior with high-frequency expressiveness in Fig. 3, reducing reliance on optical flow while improving fine-grained motion reconstruction.

Given a video sequence $\mathcal{V} = \{I_t\}_{t=1}^{T}$, each frame index t is normalized into the continuous interval $[0,1]$:

$$\tilde{t} = \frac{t-1}{T-1}, \quad \tilde{t} \in [0,1], \tag{32}$$

so that the model treats time as a continuous variable. This normalized timestamp serves as the fundamental temporal coordinate for implicit modeling, ensuring consistency across sequences of varying lengths.

To embed temporal dynamics, we adopt a sinusoidal representation network (SIREN), which is suited for encoding both smooth variations and fine high-frequency changes. The sinusoidal nonlinearity is applied in the temporal embedding $\Phi(\tilde{t})$ without replacing the activations, so it does not globally increase the model tendency to fit high-frequency noise. Each timestamp is mapped into a high-dimensional latent vector:

$$e_t = \Phi(\tilde{t}) = W_2\,\sin\big(W_1\tilde{t} + b_1\big) + b_2, \tag{33}$$

where W_1, W_2, b_1, b_2 are learnable parameters and $\sin(\cdot)$ serves as the nonlinear activation. Compared to ReLU-based mappings, this formulation provides a richer spectral fitting, enabling the model to naturally interpolate continuous dynamics and capture subtle frame-to-frame variations. To further strengthen the temporal expressiveness, we also allow multi-scale sinusoidal components to be stacked in matrix form, i.e.,

$$E_t = \big[\sin(\omega_1\tilde{t}), \cos(\omega_1\tilde{t}),\ \ldots\ , \sin(\omega_K\tilde{t}), \cos(\omega_K\tilde{t})\big], \tag{34}$$

where $\{\omega_k\}_{k=1}^{K}$ are learnable or fixed frequency coefficients. This compact representation provides a scalable way to capture temporal changes across multiple frequency bands.

The temporal embedding is broadcast across spatial dimensions and concatenated with the encoder feature map as:

$$\mathbf{H}_t(x,y) = \big[\mathbf{z}_\tau(x,y),\ e_t\big]. \tag{35}$$

In this way, every spatial position (x,y) explicitly incorporates global temporal information, providing consistent cross-frame reasoning even under noisy or misaligned conditions.

From a functional perspective, the temporal embedding can be interpreted as a continuous implicit field:

$$\phi : \mathbb{R} \to \mathbb{R}^d, \quad e_t = \phi(\tilde{t}), \tag{36}$$

which generalizes beyond discrete timestamps. Since ϕ is differentiable, temporal derivatives can also be computed,

$$v_t = \nabla_t \phi(\tilde{t}), \tag{37}$$

offering a velocity-like descriptor that implicitly encodes motion tendencies. This property enables the model to recover fine-grained temporal dynamics and preserve high-frequency details without relying on explicit pixel-level flow estimation.

To ensure consistency across long temporal ranges, ITE is applied to both forward and backward propagated features, producing enriched states $\mathbf{H}_t^{\rightarrow}$ and $\mathbf{H}_t^{\leftarrow}$. Their fusion forms a global temporal context embedding as:

$$\mathbf{H}_t = \mathbf{H}_t^{\rightarrow} \oplus \mathbf{H}_t^{\leftarrow}, \tag{38}$$

By combining explicit coordinate-driven embeddings with bidirectional propagation, the model achieves robust temporal alignment and smooth cross-frame consistency.

D. Time-Aware Spatial Graph Module

TSGM fuses the current frame with its temporal neighbors to align fine spatial details while preserving cross-frame coherence. As illustrated in Fig. 6, TSGM operates in two tightly coupled stages: time alignment, which injects continuous temporal embeddings into per-pixel features, and space alignment, which performs localized graph attention to establish reliable cross-frame correspondences. The module outputs refined per-pixel embeddings that are better suited for final reconstruction.

(i) Time alignment. We treat time as a continuous coordinate and explicitly inject a compact multi-frequency temporal code into every pixel feature. Let $\tilde{\tau} = (\tau-1)/(T-1) \in [0,1]$ be the normalized timestamp. The Fourier feature operator $\mathrm{FF}_K(\cdot)$ produces a compact multi-frequency descriptor as:

$$z_\tau = \mathrm{FF}_m(\tilde{\tau}) \in \mathbb{R}^{2m},$$

where FF_m concatenates m sine/cosine components. A small SIREN-style MLP then maps z_τ to a temporal embedding

$$\phi_\tau = \mathrm{MLP}_{\sin}(z_\tau) \in \mathbb{R}^{d_t}.$$

We broadcast and inject ϕ_τ into the encoder hidden map $\mathbf{H}_\tau$ via a learned linear mapper $\mathbf{B} \in \mathbb{R}^{d \times d_t}$ as:

$$\mathbf{U}_\tau(x,y) \;=\; \mathbf{H}_\tau(x,y) + \mathbf{B}\,\phi_\tau. \tag{39}$$

This additive injection provides an explicit continuous temporal coordinate at every spatial location, yielding *soft* cross-frame alignment without explicit optical-flow warping.

To further disambiguate motion under heavy noise, the temporal embedding is differentiable and can provide velocity-like cues. In practice, we optionally compute a finite-difference approximation of the temporal derivative,

$$\mathbf{v}_\tau \approx \frac{\phi_{\tau+1} - \phi_{\tau-1}}{2\Delta\tilde{\tau}},$$

and concatenate or linearly fuse $\mathbf{v}_\tau$ with ϕ_τ before injection in Eq. (39) when higher motion sensitivity is required.

Time alignment yields a set of time-aware feature maps $\{\mathbf{U}_\tau\}_{\tau=t-K}^{t+K}$ for subsequent spatial processing. We also compute scalar time-compatibility scores between the target timestep t and its neighbors using the ITE embeddings:

$$s(\tau) \;=\; \frac{\langle \phi_t, \phi_\tau \rangle}{\tau_{\mathrm{temp}}}, \quad \alpha_\tau \;=\; \frac{\exp(s(\tau))}{\sum_{\mu \neq t} \exp(s(\mu))}, \tag{40}$$

where τ_{temp} is a learnable temperature. The weights α_τ act as *time-aware gating* for later spatial aggregation.

(ii) Space alignment. Based on the time-aware feature maps $\{\mathbf{U}_\tau\}_{\tau=t-K}^{t+K}$, space alignment finds reliable spatial correspondences and aggregates evidence across temporally compatible frames. Nodes are tokens within each $w_s \times w_s$ local window and directed edges $t \leftarrow \tau$ are defined by the attention weights. We build a localized spatiotemporal graph by operating inside non-overlapping windows and perform cross-frame attention.

We first partition each feature map into non-overlapping windows of size $ws \times ws$; let $\mathcal{W}_{ws}(\cdot)$ denote this windowing operator and $\mathcal{W}_{ws}^{-1}(\cdot)$ its inverse that reassembles windows back to the full lattice. For window index w, we define:

$$\mathbf{U}_\tau^w \;=\; \mathcal{W}_{ws}(\mathbf{U}_\tau) \in \mathbb{R}^{N \times d}, \quad N = ws^2, \tag{41}$$

where each row of $\mathbf{U}_\tau^w$ corresponds to a node in the window.

Within each window we compute linear projections to obtain query, key, and value matrices:

$$\mathbf{Q}_t^w = \mathbf{U}_t^w \mathbf{W}_q, \quad \mathbf{K}_\tau^w = \mathbf{U}_\tau^w \mathbf{W}_k, \quad \mathbf{V}_\tau^w = \mathbf{U}_\tau^w \mathbf{W}_v,$$

where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d \times d_q}$ and $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$ reduce dimensionality for efficiency. The neighbor-to-target attention within window w is computed as:

$$\mathbf{A}_{t\leftarrow\tau}^w = \mathrm{softmax}\Big(\frac{\mathbf{Q}_t^w \mathbf{K}_\tau^{w\top}}{\sqrt{d_q}} + \beta\, \mathbf{M}_{t\leftarrow\tau}^w \Big), \tag{42}$$

$$\mathbf{O}_{t\leftarrow\tau}^w = \mathbf{A}_{t\leftarrow\tau}^w \mathbf{V}_\tau^w, \tag{43}$$

where $\mathbf{M}_{t\leftarrow\tau}^w$ is an optional logit bias derived from appearance similarity or a learned mask, and β controls its influence. Temporal aggregation then combines contributions from multiple neighbor frames using the time-gating weights α_τ in Eq. (40):

$$\mathbf{O}_t^w \;=\; \sum_\tau \alpha_\tau\, \mathbf{O}_{t\leftarrow\tau}^w, \quad \tau \in [t-K, t+K]. \tag{44}$$

The scalar α_τ favors temporally compatible frames and suppresses noisy or distant ones before spatial aggregation.

The windowed graph attention thus behaves as a localized graph neural network, where each window forms a subgraph of N nodes, with the softmaxed QK product defining adjacency and multiplication with V performing message aggregation. The full-map response from all windows is reassembled as

$$\widetilde{\mathbf{O}}_t \;=\; \mathcal{W}_{ws}^{-1}\big(\{\mathbf{O}_t^w\}_w\big). \tag{45}$$

We refine $\widetilde{\mathbf{O}}_t$ with a shallow projection block to obtain the aligned representation $\mathbf{z}_t^* = G_{\mathrm{align}}(\widetilde{\mathbf{O}}_t)$, and decode it with a convolutional mapper F_Θ, yielding the final clean prediction:

$$\widehat{C}_t = F_\Theta\big(G_{\mathrm{align}}(\widetilde{\mathbf{O}}_t)\big). \tag{46}$$

Overall, TSGM replaces explicit optical-flow warping with continuous time encoding and localized cross-frame graph attention, enabling robust spatiotemporal alignment under noise and motion. It thus provides motion-aware and coherence-preserving features for the final implicit reconstruction.

V. Experimental Results

A. Datasets and Metrics

Datasets. For comprehensive assessment across diverse conditions, we evaluate on both synthetic and real noisy datasets, covering a wide range of noise levels and motion patterns.

1) Synthetic Noise. To ensure consistency with previous works, we employ the commonly used DAVIS [46] and Set8 [1] datasets. DAVIS contains 150 high-quality natural videos with diverse object motions and complex dynamics, while Set8 consists of 8 high frame-rate sequences widely adopted in video restoration tasks. To simulate controlled degradations with varying noise levels, we add i.i.d. additive white Gaussian noise with standard deviation $\sigma \in [5, 55]$ to the clean frames.

2) Real-world Natural Noise. For real raw noise, we adopt the CRVD dataset [24], which contains 55 dynamic scenes captured under five ISO levels (1600–25600). Each indoor sequence provides multiple noise realizations with paired clean reference frames obtained by averaging repeated captures, while outdoor videos contain only noisy measurements. This dataset over 300 sequences, reflects the practical challenges of burst noise, spatial inhomogeneity, and high-ISO acquisition, serving as a reliable benchmark for real-world video denoising.

3) Microscopy Noise. To assess real-world robustness in scientific imaging, we further use public fluorescence microscopy sequences from DeepCAD [3], [36], including mouse brain neutrophils, zebrafish optic tectum neurons, and multiple zebrafish brain regions. The in vivo recordings are acquired under photon-limited conditions in challenging scenarios.

Metrics. The peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are reported to measure distortion and structural fidelity. For perceptual evaluation, we additionally use learned perceptual image patch similarity (LPIPS) for perceptual similarity, natural image quality evaluator (NIQE) for no-reference image quality, and a learned flicker detector (LFD) to quantify temporal flicker artifacts.

B. Implementation Details

For synthetic benchmarks, input sequences are cropped into 96×96 patches with temporal length $T = 5$, while for real raw-noise datasets we use $T = 3$ neighboring frames, and all frames are normalized to $[0, 1]$ before training. On synthetic data, we use a supervised negative log-likelihood loss $L_{\log}$, whereas on real noisy videos without ground truth we adopt a self-supervised blind-spot L_2 loss combined with a distillation term weighted by $\alpha = 5 \times 10^4$ to stabilize optimization. Unless otherwise stated, we optimize SINF per dataset and per sequence for real noise with a shared default configuration and no noise-type parameters; cross-dataset differences are limited to engineering settings such as T and patch/tiling for memory. We optimize with Adam [53] using an initial learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, gradient clipping with max norm 5, and weight decay 1×10^{-6}; the batch size is set to 8 for synthetic datasets and 4 for real-noise datasets. For synthetic benchmarks, we apply Gaussian noise with $\sigma \in [5, 55]$, while for real data we randomly sample across ISO levels to improve robustness. The encoder–decoder backbone uses 5 stages with channel sizes $32, 64, 128, 128, 128$ and residual blocks; the implicit temporal embedding is implemented as a 3-layer sinusoidal MLP with hidden dimension 64; the time-aware spatial graph module operates on 7×7 windows with temporal radius $K = 3$ and 4 attention heads; and the final implicit neural field head F_Θ is a 4-layer MLP with hidden width 128 and ReLU activations. All models are implemented in PyTorch and trained on a single NVIDIA RTX 4090H GPU.

C. Comparison with State-of-the-arts

To comprehensively evaluate our method, we compare SINF against a broad set of baselines, including the traditional non-learning method VBM4D [47], supervised models NAFNet [48], FastDVDnet [13], PaCNet [49], FloRNN [27], and RViDeNet [24], as well as unsupervised approaches MF2F [11], RFR [31], UDVD [5], RDRF [6], ER2R [12], TAP [6], and STBN [14]. For synthetic Gaussian denoising, MF2F, RFR, and ER2R are trained and evaluated on Set8 and DAVIS, whereas UDVD and RDRF are trained on the DAVIS training split and tested on its validation/test sets. For real raw video denoising, all unsupervised methods are trained on the CRVD training split covering both indoor and outdoor scenes and evaluated on the official indoor and outdoor test sets following the standard protocol. All competing methods are run under the same noise settings and experimental configurations, and their results are obtained either from official implementations or from the original publications.

1) Synthetic Noise: We first evaluate the proposed method on synthetic Gaussian noise following the standard settings in prior work [5], [14], [27]. Experiments are conducted on the DAVIS and Set8 datasets, which cover diverse motion patterns and scene complexities, and all models are trained under identical noise distributions for fair comparison. The quantitative results are reported in Table I. SINF consistently outperforms all unsupervised baselines, exceeding the second-best unsupervised method by more than 0.8 dB on DAVIS and over 1.1 dB on Set8 in average PSNR, with similar gains in SSIM, which highlights the effectiveness of the proposed continuous spatiotemporal modeling. Notably, SINF also achieves performance competitive with supervised methods such as FloRNN [27]. Representative visual comparisons in Fig. 7 further highlight these gains. By parameterizing video as a continuous spatiotemporal manifold and casting denoising as a coordinate-based reconstruction problem, SINF better preserves high-frequency structures and maintains temporal coherence than discrete frame-based networks that depend on local motion compensation. Competing methods such as RFR and UDVD often yield oversmoothed textures and noticeable flicker in fast-motion regions, whereas SINF recovers sharper edges, finer details, and more consistent motion patterns, producing visually cleaner and temporally more stable sequences.

2) Real-world Raw Noise: To further assess robustness under realistic noise, we evaluate SINF on the CRVD real raw video denoising benchmark, which contains sequences captured under varying ISO levels and illumination conditions. Following the standard unsupervised protocol [5], [6], all models are trained on the CRVD indoor subset and tested on both indoor and outdoor scenes. Competing approaches include the traditional supervised methods FastDVDnet [13], RViDeNet [24], MaskDnGAN [52], and FloRNN [27], as well as five unsupervised methods. As reported in Table II, our SINF achieves the best performance among all unsupervised methods at every ISO level, with an average PSNR gain of about 1.19 dB over the second-best STBN. Remarkably, SINF attains performance comparable to or better than supervised models trained with clean references, indicating strong generalization to unseen, scene-dependent noise distributions. This gain is largely attributable to the coordinate-based continuous representation, which allows SINF to implicitly learn both noise statistics and motion priors without explicit supervision.

Representative visual comparisons for indoor scenes are shown in Fig. 8, where supervised baselines such as FloRNN tend to oversmooth high-ISO regions and wash out fine textures, while UDVD and other unsupervised methods often leave residual high-frequency noise or mild temporal flicker. In these settings, SINF better suppresses noise while preserving fine structures on walls, objects, and textures, leading to cleaner yet detailed reconstructions. Besides, Fig. 9 further reports results on outdoor sequences with stronger motion and illumination variations, where existing denoisers degrade more noticeably, either blurring moving objects or amplifying flicker under changing lighting. In contrast, SINF maintains clearer boundaries, more consistent appearance across frames, and sharper details in both foreground and background, yielding more visually faithful reconstructions in real-world conditions.

3) Microscopy Noise.: The evaluation on scientific imaging is conducted on the public fluorescence microscopy data from DeepCAD [36], covering both synthetic two-photon calcium

TABLE I: Quantitative evaluation of video denoising results, comparing to the representative methods that are traditional, supervised or unsupervised approaches. The best and second results among unsupervised methods are **highlighted** and underline.

		Method	$\sigma=10$		$\sigma=20$		$\sigma=30$		$\sigma=40$		$\sigma=50$		Average	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DAVIS	Traditional	VBM4D [47]	37.58	-	33.88	-	31.65	-	30.05	-	28.80	-	32.39	-
	Supervised	NAFNet [48]	38.79	0.965	35.37	0.933	33.47	0.904	32.17	0.879	31.18	0.858	34.20	0.908
		FastDVDNet [13]	38.71	0.962	35.77	0.941	34.04	0.917	32.82	0.895	31.86	0.875	34.64	0.919
		PaCNet [49]	39.97	0.971	37.10	0.947	35.07	0.921	33.57	0.897	32.39	0.874	35.62	0.922
		FloRNN [27]	40.16	0.976	37.52	0.956	35.89	0.944	34.66	0.929	33.67	0.913	36.38	0.944
	Unsupervised	MF2F [11]	38.04	0.957	35.61	0.936	33.65	0.907	31.50	0.852	29.39	0.784	33.64	0.887
		RFR [50]	39.31	0.969	36.15	0.942	34.28	0.917	32.92	0.893	31.86	0.873	34.90	0.918
		UDVD [5]	39.17	0.970	35.94	0.943	34.09	0.918	32.79	0.895	31.80	0.874	34.76	0.920
		RDRF [51]	39.54	0.972	36.40	0.947	34.55	0.925	33.23	0.903	32.20	0.883	35.18	0.926
		ER2R [12]	39.52	-	36.49	-	34.60	-	33.29	-	32.25	-	35.23	-
		TAP [6]	39.69	0.972	36.62	0.948	34.71	0.925	33.37	0.903	32.36	0.884	35.35	0.926
		STBN [14]	40.35	0.961	37.67	0.961	36.00	0.945	34.73	0.930	33.70	0.914	36.49	0.945
		SINF (Ours)	**41.38**	**0.983**	**38.19**	**0.979**	**36.77**	**0.958**	**35.69**	**0.956**	**34.83**	**0.912**	**37.37**	**0.958**
Set8	Traditional	VBM4D [47]	36.05	-	32.19	-	30.00	-	28.48	-	27.33	-	30.81	-
	Supervised	NAFNet [48]	36.52	0.943	33.55	0.802	31.81	0.869	30.59	0.842	29.65	0.818	32.43	0.875
		FastDVDNet [13]	36.44	0.954	33.43	0.920	31.68	0.889	30.46	0.861	29.53	0.835	32.31	0.892
		PaCNet [49]	37.06	0.960	33.94	0.925	32.05	0.892	30.70	0.862	29.66	0.835	32.68	0.895
		FloRNN [27]	37.57	0.964	34.67	0.938	32.97	0.914	31.75	0.891	30.80	0.870	33.55	0.915
	Unsupervised	MF2F [11]	36.01	0.938	33.79	0.912	32.20	0.883	30.64	0.841	28.90	0.778	32.31	0.870
		RFR [50]	36.77	0.953	33.64	0.922	31.82	0.886	30.52	0.864	29.50	0.828	32.45	0.891
		UDVD [5]	36.36	0.951	33.53	0.917	31.88	0.887	30.72	0.859	29.81	0.835	32.46	0.890
		RDRF [51]	36.67	0.955	34.00	0.925	32.39	0.898	31.23	0.873	30.31	0.850	32.92	0.900
		ER2R [12]	37.55	-	34.34	-	32.45	-	31.09	-	30.05	-	33.10	-
		TAP [6]	38.02	0.958	35.07	0.927	33.42	0.900	32.10	0.875	31.16	0.852	33.95	0.902
		STBN [14]	37.24	0.959	34.41	0.932	32.76	0.907	31.57	0.884	30.62	0.861	33.32	0.909
		SINF (Ours)	**38.79**	**0.975**	**36.10**	**0.948**	**34.55**	**0.928**	**33.37**	**0.899**	**32.69**	**0.887**	**35.10**	**0.927**

TABLE II: Quantitative evaluation of raw video denoising results, comparing to the representative methods that are supervised and unsupervised approaches. The best and second results among unsupervised methods are **highlighted** and underline.

		Method	ISO=1600		ISO=3200		ISO=6400		ISO=12800		ISO=25600		Average	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CRVD	Supervised	FastDVDNet [13]	43.43	0.987	42.91	0.984	40.29	0.979	36.05	0.961	36.50	0.940	39.84	0.970
		RViDeNet [24]	47.74	0.994	45.91	0.991	43.85	0.988	41.20	0.982	41.17	0.982	43.97	0.987
		MaskDnGAN [52]	47.52	0.994	45.88	0.991	44.14	0.987	41.48	0.983	40.79	0.982	43.96	0.989
		FloRNN [27]	48.81	0.996	47.05	0.993	45.09	0.991	42.63	0.987	42.19	0.987	45.15	0.991
	Unsupervised	UDVD [5]	48.02	0.998	46.44	0.998	44.74	0.997	42.21	0.997	42.13	0.995	44.71	0.997
		RDRF [51]	48.38	0.998	46.86	0.998	45.24	0.998	42.72	0.997	42.25	0.995	45.09	0.997
		ER2R [12]	49.14	-	47.51	-	45.61	-	43.03	-	42.91	-	45.64	-
		TAP [6]	48.85	0.992	47.03	0.899	45.11	0.991	42.44	0.987	42.33	0.986	45.15	0.995
		STBN [14]	49.27	**0.999**	47.58	0.999	45.75	0.998	43.36	**0.998**	42.91	0.997	45.77	0.998
		SINF (Ours)	**50.43**	0.998	**48.47**	**0.999**	**46.93**	**0.999**	**44.75**	0.995	**44.24**	**0.998**	**46.96**	**0.998**

imaging and in vivo sequences of mouse brain neutrophils, zebrafish optic tectum neurons, and multiple zebrafish brain regions. For the synthetic setting, different input SNR levels are generated by varying the relative photon counts, and the results of DeepSeMi [37], DeepCAD [36], SRDTrans [38], and SINF are reported in Table III. Across all five input SNR conditions, SINF consistently achieves the highest reconstruction SNR and the best average score, with particularly clear gains in the most photon-starved regime. For real fluorescence sequences, we temporally subsample the videos to induce faster apparent motion and stronger frame-to-frame intensity fluctuations, and visually shown in Fig. 10. SINF more strongly suppresses photon-limited noise while better preserving sharp textures and delicate structures, yielding reconstructions that are visually cleaner and more faithful to the underlying morphology.

D. Robustness and Generalization

We assess model robustness by systematically varying both noise type and strength. Figs. 11a–11c present PSNR curves under additive Gaussian noise, Poisson noise, and mixed Gaussian–Poisson degradations on the DAVIS dataset. Across all three settings, existing unsupervised baselines, including UDVD, STBN, and TAP, exhibit steeper performance drops and larger dispersion as noise intensity increases. SINF consistently attains the highest PSNR with the smallest variance, indicating stable behavior even under severe corruption. Temporal robustness is then evaluated on CRVD high-ISO sequences with strong real-world degradations. In the frame-drop experiment shown in Fig. 11d, consecutive frames are randomly removed to emulate missing or corrupted inputs, our SINF shows the most gradual PSNR decay as the drop ratio grows. In the frame-jitter experiment in Fig. 11e, temporal indices are perturbed to mimic camera shake; SINF still produces stable, temporally consistent reconstructions, while competing methods exhibit a performance drop. These results demonstrate that the continuous spatiotemporal representation yields pronounced robustness to both spatial and temporal degradations. To examine generalization, we perform few-shot adaptation on CRVD under real noise. Models are first trained on synthetic data and then adapted to CRVD using only a small subset of target-domain frames. As illustrated in Fig. 11f,

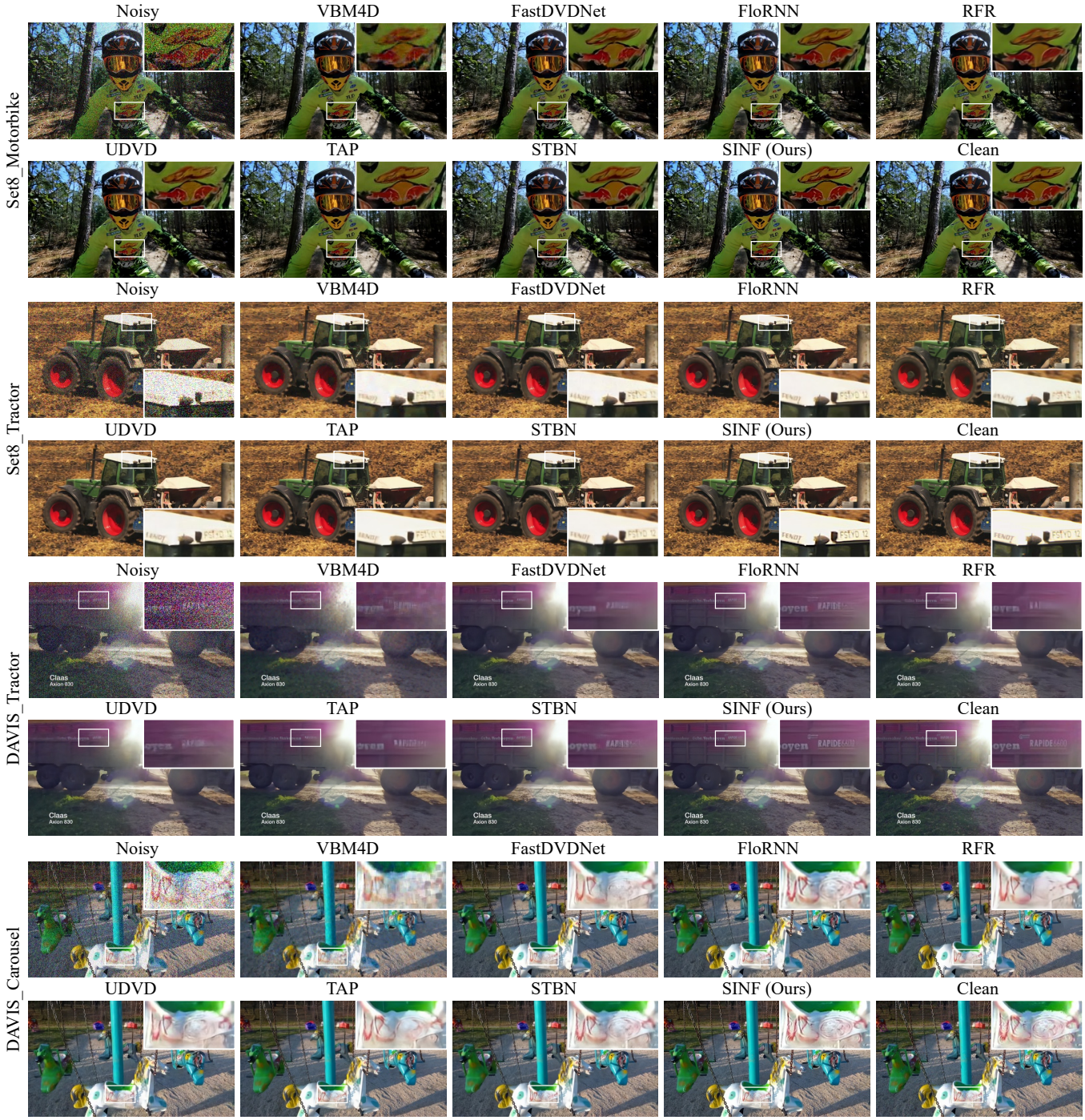


Fig. 7: Qualitative comparison on synthetic Gaussian noise from DAVIS and Set8 datasets.

SINF improves steadily as the adaptation ratio increases and already achieves strong performance with as little as 1–5% adaptation data. Compared with other unsupervised methods, SINF offers higher zero-shot performance and larger gains in the low-data regime, reflecting better transferability and data efficiency. These experiments demonstrate that SINF maintains strong robustness under diverse noise and temporal perturbations, while its continuous spatiotemporal neural field further promotes coherent cross-domain transfer across degradation types by modeling videos as a unified signal manifold rather than dataset-specific discrete mappings, thereby supporting practical denoising under mismatched real-world conditions.

E. Ablation Study

1) Blind-Spot Implicit Field Modeling: To evaluate the effectiveness of the proposed BS-INF module, we replaced the implicit field component with three representative blind-spot mechanisms, namely rotational, shuffled, and bidirectional blind-spot aggregation adopted from UDVD [5], PUCA [54], and STBN [14], respectively. The ablation experiments were conducted on the Set8 dataset under Gaussian noise with $\sigma = 25$. As shown in Fig. 12a, BS-INF exhibits significantly faster and smoother convergence compared with the convolution-based variants, while achieving consistently higher PSNR with the smallest parameter footprint (less than 1M). This

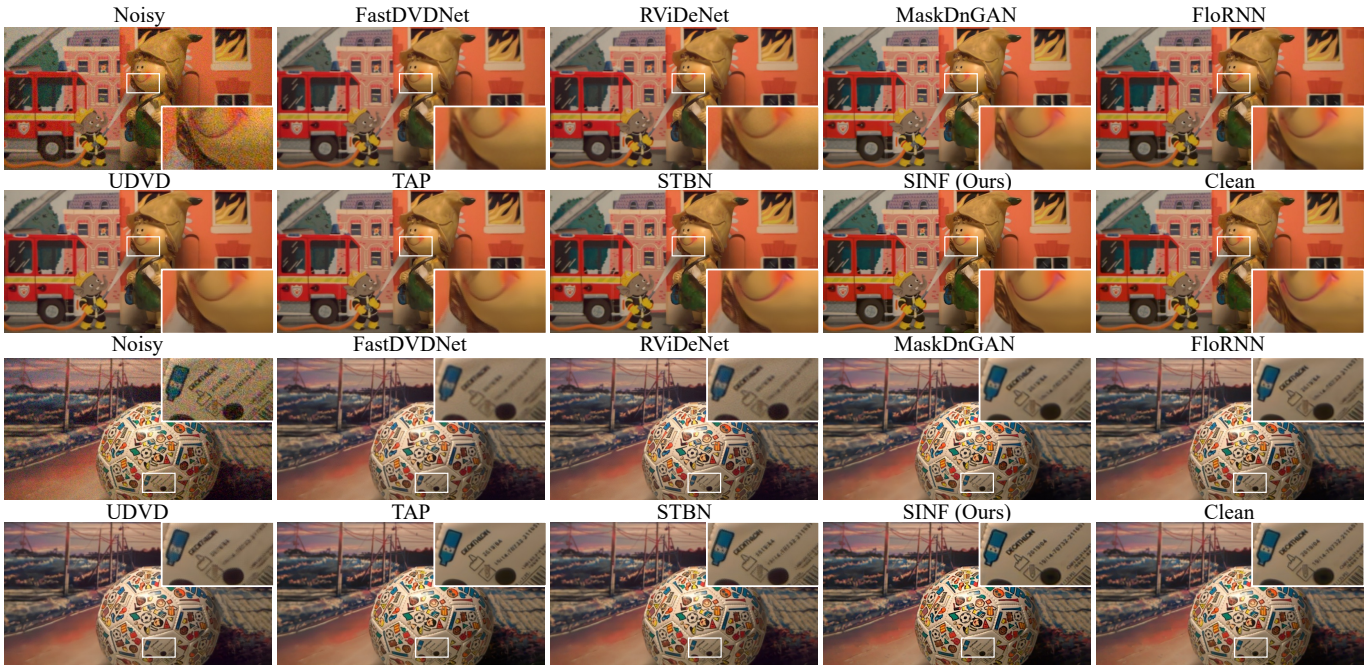


Fig. 8: Visual comparison on CRVD indoor scenes. SINF better suppresses noise while preserving more structural details.

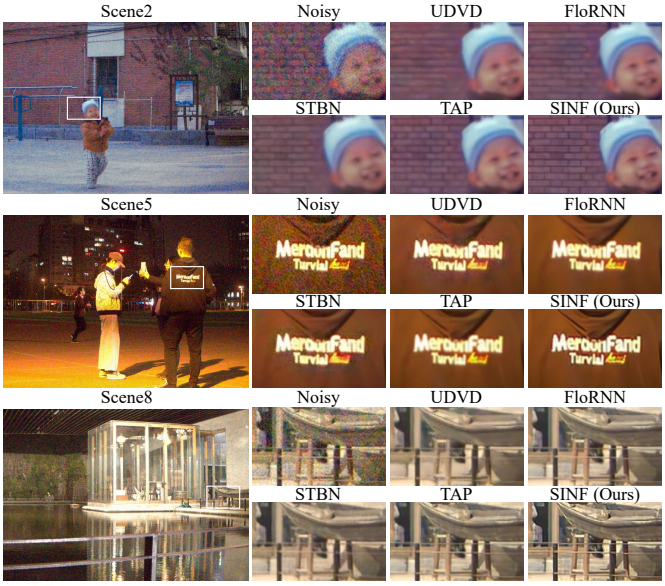


Fig. 9: Visual comparison on CRVD outdoor scenes with real-world strong motion and illumination changes.

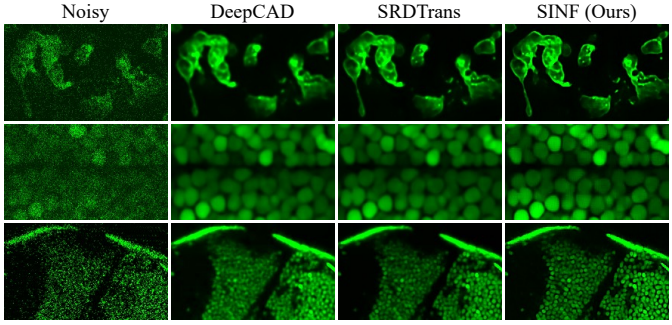


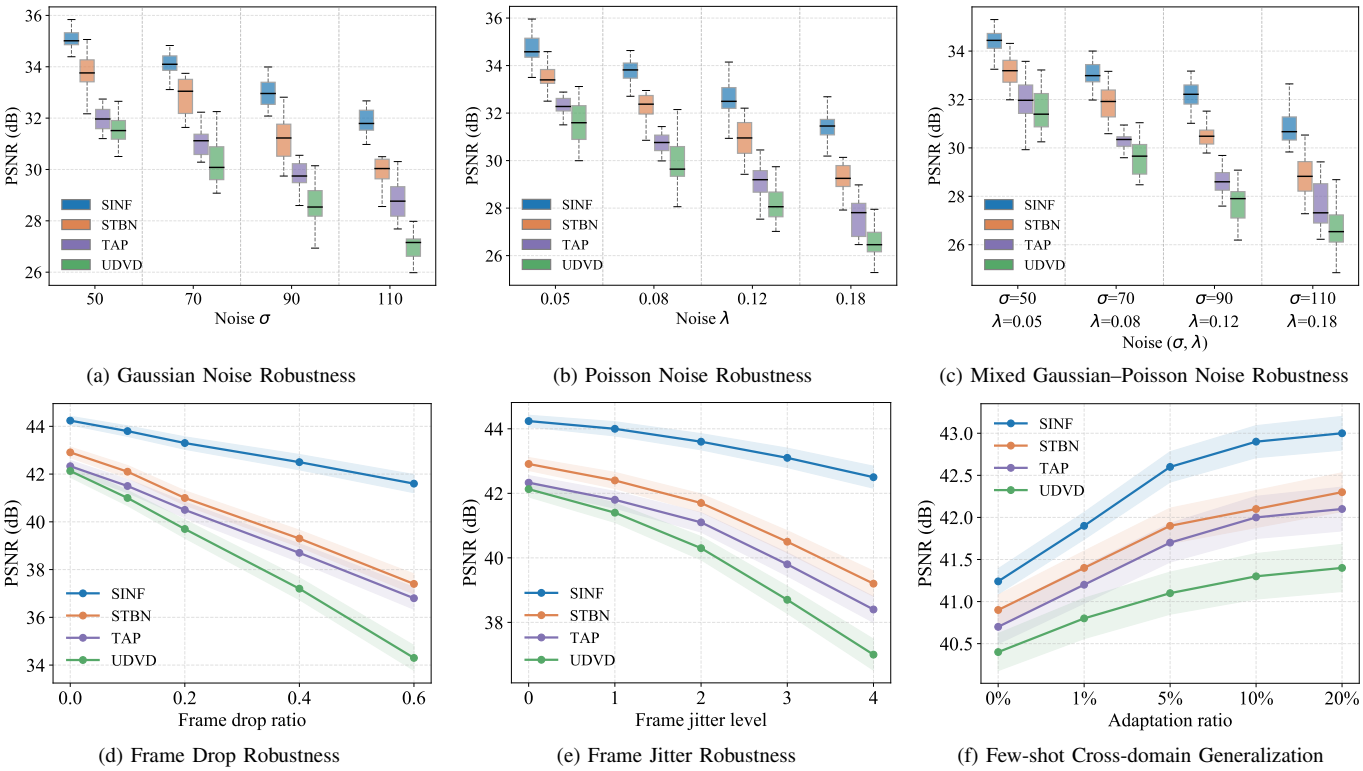
Fig. 10: Denoising results on publicly available real-world in vivo mouse and zebrafish fluorescence microscopy datasets.

Method	SNR (dB) at different input levels					Avg.
	-5.52	-2.51	0.48	2.88	5.17	
DeepSeMi [37]	20.80	23.10	24.50	25.70	26.60	24.14
DeepCAD [36]	21.20	23.65	25.00	26.30	27.20	24.67
SRDTrans [38]	21.90	24.10	25.70	27.00	28.10	25.36
SINF (Ours)	**22.60**	**24.80**	**26.60**	**28.00**	**29.10**	**26.22**

TABLE III: Reconstruction analysis on synthetic two-photon calcium imaging sequences under different input SNR levels.

demonstrates that spatial implicit field modeling not only enhances optimization stability but also achieves more efficient representation of spatial dependencies. Furthermore, Fig. 12b presents the frequency-domain energy consistency of restored images compared with GT. BS-INF maintains a frequency distribution that closely follows the ground truth spectrum, especially in the high-frequency region, where conventional blind-spot convolutions tend to lose fine structural information. These results suggest that the spatial implicit field leverages global coordinate priors to preserve high-frequency structure, leading to more faithful spatial reconstruction under noise.

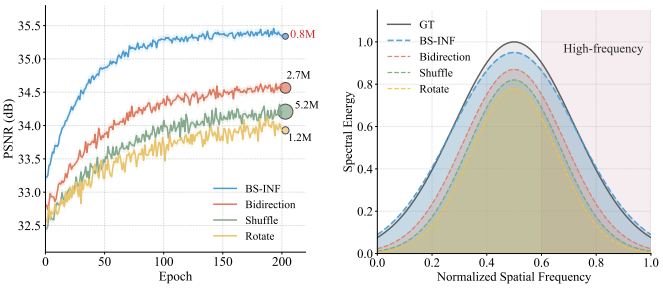
2) Temporal Alignment Validation: To validate the effectiveness of our temporal implicit modeling, we perform an ablation under challenging motion with high-speed object movement and strong additive noise. We compare three temporal alignment strategies on a synthetic Set8 sequence with large motion and Gaussian noise $\sigma = 70$, and on a real CRVD sequence with camera shake: a variant without any temporal alignment, a variant with explicit optical-flow warping, and our full model with ITE operator. Table IV reports PSNR, SSIM, and temporal consistency error (TC), where TC is defined as the average ℓ_1 difference between each frame and its adjacent frame warped by the estimated flow; lower TC indicates smoother temporal evolution. The ITE variant consistently achieves the highest PSNR and SSIM and the lowest TC on both datasets, showing that the implicit temporal field better preserves both coherence and fidelity. To make the behavior under strong noise and large motion more



(a) Gaussian Noise Robustness (b) Poisson Noise Robustness (c) Mixed Gaussian–Poisson Noise Robustness

(d) Frame Drop Robustness (e) Frame Jitter Robustness (f) Few-shot Cross-domain Generalization

Fig. 11: Comprehensive evaluation of different methods across diverse degradation and adaptation scenarios. (a–c) Noise robustness under Gaussian, Poisson, and mixed Gaussian–Poisson degradations. (d–e) Temporal robustness against frame drop and frame jitter distortions. (f) Cross-domain generalization under limited adaptation frames.



(a) Training Stability (b) Spectral Energy Distribution

Fig. 12: Model analysis of different spatial blind-spot designs. (a) Training convergence and parameter count. (b) Frequency spectra consistency of restored images compared with GT.

TABLE IV: Quantitative comparison of temporal alignment strategies, evaluated by PSNR/SSIM and temporal consistency.

Category	Method	PSNR↑ (dB) / SSIM↑		TC↓
		Set8	CRVD	
w/o Flow	FastDVDnet	33.0 / 0.915	31.2 / 0.893	0.142
	UDVD	32.6 / 0.910	30.8 / 0.881	0.151
	FloRNN	33.3 / 0.918	31.5 / 0.895	0.136
with Flow	MF2F	33.5 / 0.920	31.7 / 0.897	0.122
	RDRF	<u>33.8</u> / <u>0.923</u>	31.9 / <u>0.902</u>	<u>0.118</u>
	STBN	34.1 / 0.928	<u>32.3</u> / 0.907	0.107
with ITE	**SINF (Ours)**	**34.7 / 0.936**	**33.0 / 0.915**	**0.082**

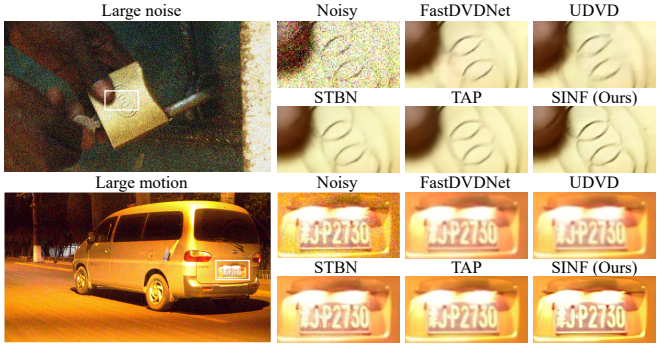


Fig. 13: Visual comparisons under strong noise and large motion on Set8 with σ=70 and CRVD with camera shake.

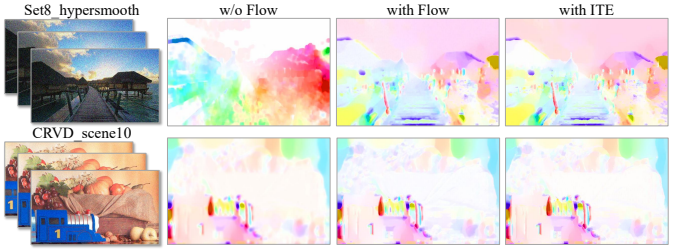
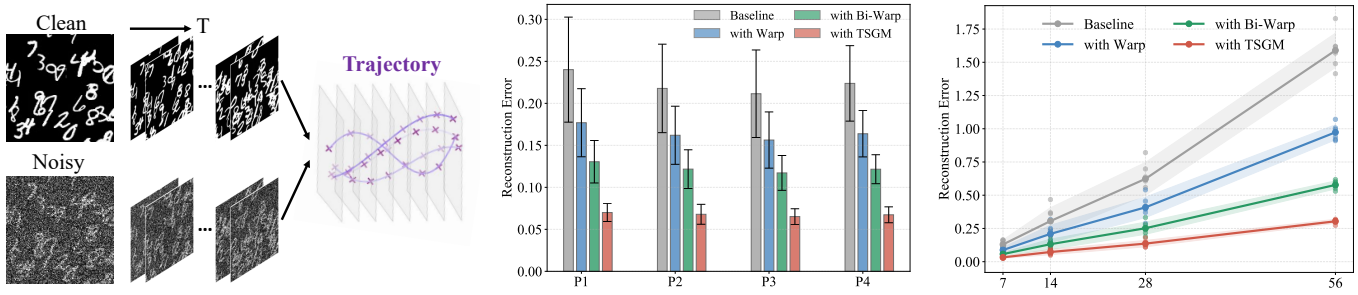


Fig. 14: Comparative visualization of temporal alignment strategies, including w/o Flow, with explicit Flow, and with our ITE, on both synthetic and real noisy video sequences.

directly verifiable and better emulate harsh real-world imaging conditions, Fig. 13 shows representative crops under the same protocol, where our ITE-based modeling yields sharper and cleaner reconstructions with fewer motion-induced artifacts than explicit warping. Fig. 14 further visualizes the estimated motion fields: removing temporal alignment yields fragmented and incoherent motion, while explicit flow warping improves local alignment but still introduces discontinuities and warping



(a) Tracked pixel trajectories. (b) Reconstruction consistency. (c) Long-sequence temporal consistency.

Fig. 15: Pixel tracking and reconstruction consistency across temporal alignment strategies. (a) Tracked pixel trajectories. (b) Reconstruction error and temporal variance across seed pixels. (c) Long-sequence robustness as the sequence length increases.

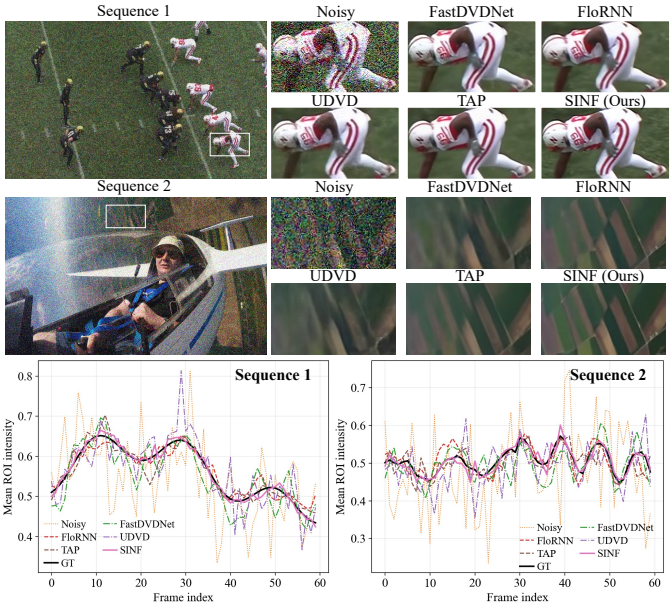


Fig. 16: Spatiotemporal consistency analysis of long video sequences. Visual comparison of denoised frames and the corresponding ROI intensity trajectories across different methods.

artifacts. In contrast, ITE produces globally coherent motion patterns and stable reconstructions even under severe noise and large displacements, indicating a more robust temporal alignment mechanism than explicit flow.

3) Time-Aware Spatial Graph Module: For further analysis of the spatiotemporal stability brought about by TSGM, we conducted a pixel-tracking experiment on a controlled motion sequence, where four pixels were randomly selected and tracked across frames. The trajectories in Fig. 15a reveal that the Baseline and the single-warp variant (*with Warp*) suffer from noticeable drift and jitter, while the bi-directional warp (*with Bi-Warp*) reduces drift but still accumulates errors due to flow estimation uncertainty. In contrast, TSGM produces smooth and stable tracks with the most consistent temporal correspondence. The reconstruction error statistics in Fig. 15b show that TSGM attains the lowest temporal variance across all seed pixels, indicating reduced flicker and improved temporal consistency. Furthermore, the robustness study in Fig. 15c, where the sequence length is varied as $T \in \{7, 14, 28, 56\}$, demonstrates that competing strategies degrade as T increases,

TABLE V: Performance comparison of different graph attention modules. Best and second-best are **bold** and underlined.

#	Modules			PSNR↑	
	+GNN	+Time Align.	+Space Align.	DAVIS	CRVD
A	✗	✗	✗	34.35	32.16
B	✓	✗	✗	34.52	32.30
C	✗	✓	✗	<u>34.98</u>	32.65
D	✗	✗	✓	34.84	<u>32.71</u>
E	✗	✓	✓	**35.42**	**33.16**

whereas TSGM maintains a significantly lower error curve and is more resilient to long-range temporal drift. This varying-length tracking study indicates that the bounded-window graph alignment remains stable as the sequence grows longer, mitigating the risk of error accumulation over time. These results verify that our method effectively bridges spatial and temporal alignment, enabling stable pixel correspondence and improved restoration quality under long sequences and complex motion.

We further quantify the contribution of each component through an ablation study on DAVIS and CRVD, summarized in Table V. Starting from a baseline model, adding a spatial graph neural network (GNN) is useful, but the gain remains limited without explicit temporal modeling. Activating only the temporal alignment module stabilizes motion association over time, while activating only the spatial alignment module strengthens local feature aggregation. When temporal and spatial alignment are jointly enabled in the full TSGM configuration, the model attains the best performance on both datasets, indicating that jointly modeling temporal correspondence and spatial graph interactions is crucial for robust video restoration.

4) Spatiotemporal Trajectory Consistency: To probe spatiotemporal consistency beyond frame-wise scores, we conduct a trajectory-level analysis on two long sequences, randomly selecting one from Set8 and one from DAVIS under the same synthetic Gaussian noise setting as in the main experiments. For each sequence, we track a small region of interest (ROI) on a moving object and compute, at every frame, the mean ROI intensity, yielding 1D curves that reflect the temporal evolution of local appearance. As shown in Fig. 16, the upper panels visualize cropped regions around the tracked area, while the lower panels plot ROI intensity trajectories for the noisy input, FloRNN, FastDVDnet, UDVD, TAP, SINF, and the GT. The noisy curves show strong high-frequency fluctuations, supervised baselines such as FastDVDnet and FloRNN sup-

TABLE VI: Quantitative analysis among activation functions.

Activation	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	LFD↓
Linear	32.48	0.877	0.195	4.12	17.85
Sigmoid	32.92	0.881	0.189	4.08	17.23
Tanh	33.08	0.885	0.183	4.03	16.82
ReLU	33.25	0.889	0.176	3.98	16.43
GELU	33.47	0.892	0.168	3.91	15.87
Swish	33.82	0.896	0.161	3.78	14.95
SIREN (Ours)	**34.55**	**0.928**	**0.138**	**3.41**	**12.51**

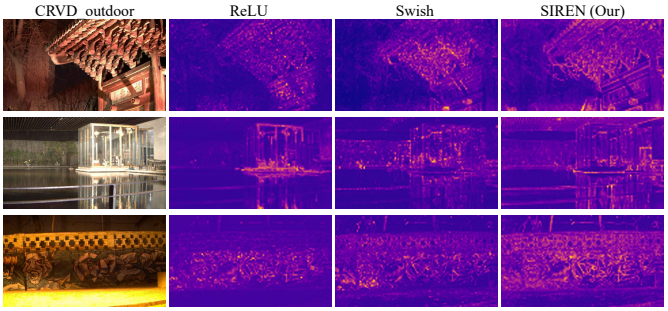


Fig. 17: Response visualization of three activation functions.

press noise but over-smooth the temporal signal and introduce visible bias, and unsupervised baselines like UDVD and TAP preserve more variation yet exhibit residual jitter and irregular oscillations indicative of temporal flicker. In contrast, the SINF trajectory closely follows the ground truth in both amplitude and phase, reducing noise while preserving the underlying temporal dynamics. This tighter curve alignment, together with sharper yet temporally stable visual details, indicates that SINF achieves more faithful spatiotemporal consistency than existing supervised and unsupervised methods.

5) Activation Function Analysis: We analyze different activation functions on representation learning by replacing the nonlinearity with Linear, Sigmoid, Tanh, ReLU, GELU, Swish, and our periodic sinusoidal activation SIREN, using Set8 under Gaussian noise with $\sigma = 30$. In addition to distortion metrics, we report LPIPS, NIQE, and LFD to assess perceptual quality. As shown in Table VI, the Linear variant performs the worst due to severely limited expressiveness. Sigmoid and Tanh improve over the linear case but suffer from saturation, compressing dynamic range and attenuating subtle intensity variations. ReLU further boosts performance via sparse activation, yet its piecewise-linear and non-periodic form still suppresses fine-scale contrast and high-frequency details. Smoother gates such as GELU and Swish achieve stronger results because their soft input-dependent responses preserve more mid-frequency content, but they still behave effectively as low-pass filters with limited spectral coverage. In contrast, the periodic and infinitely differentiable nature of SIREN endows the implicit field with richer Fourier support and stable high-order gradients, enabling continuous modeling of both low- and high-frequency components in space and time. This advantage is reflected by consistently better perceptual scores and visually sharper reconstructions. Response visualizations in Fig. 17 show that our SIREN preserves thin structures and sharp boundaries in challenging CRVD outdoor scenes, whereas ReLU- and Swish-based variants tend to oversmooth edges and wash out fine textures. Overall, periodic

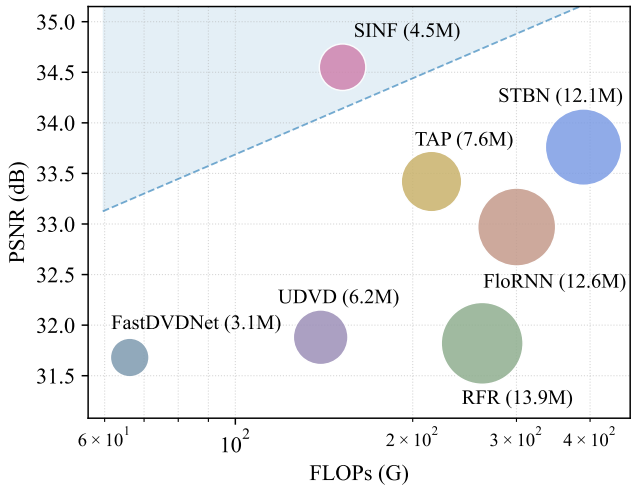


Fig. 18: Accuracy–efficiency comparison of different methods. The x-axis denotes FLOPs, the y-axis denotes PSNR, and the circle area is proportional to the number of parameters.

TABLE VII: Computational efficiency and scalability of SINF under varying temporal windows and spatial resolutions.

(a) Scaling with sequence length.

Temporal window	Time (ms/fr)↓	Mem (GB)↓	Throughput (fps)↑
1	6.85	4.2	146.0
3	8.02	7.6	124.7
5	9.13	11.5	109.5
7	10.48	13.4	95.4
9	11.96	15.1	83.6

(b) Scaling with spatial resolution.

Resolution	Time (ms/fr)↓	Mem (GB)↓	Throughput (fps)↑
256×256	9.13	11.5	109.5
320×320	13.70	14.2	73.0
384×384	19.80	17.6	50.5
448×448	27.10	21.9	36.9
512×512	35.90	27.4	27.9

continuous-field activations like SIREN provide a more powerful implicit representation for high-frequency, temporally coherent restoration than standard non-periodic nonlinearities.

6) Efficiency and Scalability: To further assess practical deployability, we report a unified efficiency profile by quantifying the efficiency and scaling behavior. Fig. 18 visualizes the accuracy–efficiency trade-off across representative baselines, including FastDVDNet [13], FloRNN [27], RFR [50], UDVD [5], TAP [6], and STBN [14]. The x-axis reports FLOPs, the y-axis reports Set8 PSNR at σ=30, and the marker area is proportional to the parameter count. SINF falls in a favorable region of the curve, indicating improved restoration quality without a disproportionate increase in computational footprint. Table VII further quantifies the runtime and memory scaling of SINF under identical implementation settings. Subtable VIIa varies the temporal window length T and reports latency, peak GPU memory, and throughput, making the cost growth with increasing temporal context explicit. Subtable VIIb varies the spatial resolution with fixed T=5 and measures how latency and memory increase as the number of queried coordinates grows. Notably, dense coordinate querying is realized as a lightweight feature-conditioned implicit component without without incurring redundant per-location

computation. These results directly quantify the deployability and scaling of SINF across resolutions and sequence lengths.

VI. Conclusion

In this paper, we presented SINF, a self-supervised video denoising framework that rethinks video as a continuous function over space and time rather than as a discretized frame grid. By introducing a blind-spot implicit spatial field, an implicit temporal embedding, and a time-aware spatial graph module, the framework unifies spatial and temporal reasoning in a coordinate-based implicit representation and directly models the noisy image formation process in a continuous spatiotemporal domain. This design breaks the receptive-field limitations of blind-spot CNNs and avoids the noise sensitivity of flow-based alignment, enabling globally informed texture reconstruction and motion-consistent denoising under severe noise and complex dynamics. Extensive experiments on synthetic and real noisy benchmarks demonstrate consistent gains in both accuracy and perceptual quality, with better high-frequency detail preservation and substantially fewer motion artifacts than prior self-supervised methods. To our knowledge, this is the first attempt to establish a spatially and temporally continuous implicit-field paradigm for self-supervised video denoising, and the proposed formulation offers a flexible foundation that can be extended to broader spatiotemporal restoration and enhancement tasks in real-world scenarios.

References

[1] M. Tassano, J. Delon, and T. Veit, "Dvdnet: A fast network for deep video denoising," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1805–1809.

[2] K. Wei, Y. Fu, J. Yang, and H. Huang, "A physics-based noise formation model for extreme low-light raw denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2758–2767.

[3] X. Li, G. Zhang, J. Wu, Y. Zhang, Z. Zhao, X. Lin, H. Qiao, H. Xie, H. Wang, L. Fang *et al.*, "Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising," *Nature methods*, vol. 18, no. 11, pp. 1395–1400, 2021.

[4] T. Liu, M. Xu, S. Li, J. Zhang, and L. Jiang, "Unifes: A unified recurrent network for quality enhancement and stabilization in face videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[5] D. Y. Sheth, S. Mohan, J. L. Vincent, R. Manzorro, P. A. Crozier, M. M. Khapra, E. P. Simoncelli, and C. Fernandez-Granda, "Unsupervised deep video denoising," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1759–1768.

[6] Z. Fu, L. Guo, C. Wang, Y. Wang, Z. Li, and B. Wen, "Temporal as a plugin: Unsupervised video denoising with pre-trained image denoisers," in *European Conference on Computer Vision*. Springer, 2024, pp. 349–367.

[7] H. Li, W. Zhang, X. Hu, T. Jiang, Z. Chen, and H. Wang, "Prompt-sid: Learning structural representation prompt via latent diffusion for single image denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4734–4742.

[8] H. Li, Y. Wang, T. Huang, H. Huang, H. Wang, and X. Chu, "Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 684–13 694.

[9] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[10] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2129–2137.

[11] V. Dewil, J. Anger, A. Davy, T. Ehret, G. Facciolo, and P. Arias, "Self-supervised training for blind multi-frame video denoising," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2724–2734.

[12] H. Zheng, T. Pang, and H. Ji, "Unsupervised deep video denoising with untrained network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3651–3659.

[13] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1354–1363.

[14] Z. Chen, T. Jiang, X. Hu, W. Zhang, H. Li, and H. Wang, "Spatiotemporal blind-spot network with calibrated flow alignment for self-supervised video denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2411–2419.

[15] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[17] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7537–7547.

[18] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2. Ieee, 2005, pp. 60–65.

[19] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[20] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[21] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[23] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[24] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, "Supervised raw video denoising with a benchmark dataset on dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2301–2310.

[25] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Transactions on Image Processing*, 2024.

[26] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool, "Recurrent video restoration transformer with guided deformable attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 378–393, 2022.

[27] J. Li, X. Wu, Z. Niu, and W. Zuo, "Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones," in *European Conference on Computer Vision*. Springer, 2022, pp. 592–609.

[28] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4947–4956.

[29] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5972–5981.

[30] L. Sun, F. Wu, W. Ding, X. Li, J. Lin, W. Dong, and G. Shi, "Multi-scale spatio-temporal memory network for lightweight video denoising," *IEEE Transactions on Image Processing*, vol. 33, pp. 5810–5823, 2024.

[31] T. Pang, H. Zheng, Y. Quan, and H. Ji, "Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2043–2052.

[32] S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[33] L. Fan, J. Cui, H. Li, X. Yan, H. Liu, and C. Zhang, "Complementary blind-spot network for self-supervised real image denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 10 107–10 120, 2024.
[34] C. Qu, Z. Chen, J. Zhang, X. Chen, and J. Han, "Self-bsr: Self-supervised image denoising and destriping based on blind-spot regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
[35] M. Claus and J. Van Gemert, "Videnn: Deep blind video denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.
[36] X. Li, Y. Li, Y. Zhou, J. Wu, Z. Zhao, J. Fan, F. Deng, Z. Wu, G. Xiao, J. He *et al.*, "Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit," *Nature biotechnology*, vol. 41, no. 2, pp. 282–292, 2023.
[37] G. Zhang, X. Li, Y. Zhang, X. Han, X. Li, J. Yu, B. Liu, J. Wu, L. Yu, and Q. Dai, "Bio-friendly long-term subcellular dynamic recording by self-supervised image enhancement microscopy," *Nature Methods*, vol. 20, no. 12, pp. 1957–1970, 2023.
[38] X. Li, X. Hu, X. Chen, J. Fan, Z. Zhao, J. Wu, H. Wang, and Q. Dai, "Spatial redundancy transformer for self-supervised fluorescence image denoising," *Nature Computational Science*, vol. 3, no. 12, pp. 1067–1080, 2023.
[39] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638.
[40] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1575–1584.
[41] C. Kim, J. Lee, and J. Shin, "Zero-shot blind image denoising via implicit neural representations," *arXiv preprint arXiv:2204.02405*, 2022.
[42] H. Li, X. Hu, and H. Wang, "Interpretable unsupervised joint denoising and enhancement for real-world low-light scenarios," in *International conference on learning representations*, 2025.
[43] X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang, "Boosting neural representations for videos with a conditional decoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2556–2566.
[44] Z. Chen, Y. Chen, J. Liu, X. Xu, V. Goel, Z. Wang, H. Shi, and X. Wang, "Videoinr: Learning video implicit neural representation for continuous space-time super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2047–2057.
[45] M. D. Aiyetigbo, W. Yuan, F. Luo, and N. Li, "Implicit neural representation for video restoration," *arXiv preprint arXiv:2506.05488*, 2025.
[46] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.
[47] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms," *IEEE Transactions on image processing*, vol. 21, no. 9, pp. 3952–3966, 2012.
[48] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European conference on computer vision*. Springer, 2022, pp. 17–33.
[49] G. Vaksman, M. Elad, and P. Milanfar, "Patch craft: Video denoising by deep modeling and patch matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2157–2166.
[50] S. Lee, D. Cho, J. Kim, and T. H. Kim, "Restore from restored: Video restoration with pseudo clean video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3537–3546.
[51] Z. Wang, Y. Zhang, D. Zhang, and Y. Fu, "Recurrent self-supervised video denoising with denser receptive field," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7363–7372.
[52] A. Paliwal, L. Zeng, and N. K. Kalantari, "Multi-stage raw video denoising with adversarial loss and gradient mask," in *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2021, pp. 1–10.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[54] H. Jang, J. Park, D. Jung, J. Lew, H. Bae, and S. Yoon, "Puca: patch-unshuffle and channel attention for enhanced self-supervised image denoising," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

Xiaowan Hu (Member, IEEE) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2019, and the Ph.D. degree in engineering from Tsinghua University, Beijing, China, in 2024. Since 2025, she has been an Assistant Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China. Her research interests include image processing, computational imaging, and computer vision. She has authored or coauthored more than 20 papers in international journals and conference proceedings, such as Nature Computational Science, IEEE Transactions on Circuits and Systems for Video Technology, Conference on CVPR, ICML, and IJCAI. Contact her at huxiaowan@buaa.edu.cn.

Henan Liu received the B.S. degree from Beihang University, Beijing, China, in 2025. Since 2025, she has been pursuing the Ph.D. degree at Beihang University, Beijing, China. Her research interests include image processing, computational imaging, and computer vision. Contact her at lhn21373089@buaa.edu.cn.

Ce Zheng is currently pursuing a Ph.D. at Beihang University. She graduated from the University of Chinese Academy of Sciences with a master's degree. Her research interests include data mining, network measurements, video compression, and large language models. Contact her at zhengce@buaa.edu.cn.

Xinyang Li is an Assistant Professor at the College of AI of Tsinghua University, Beijing, China. He received his PhD degree from Tsinghua University in 2023. His research interests include artificial intelligence, neuroscience, and optical imaging. His work has been published in top international journals, including Nature Methods, Nature Biotechnology, Nature Computational Science, etc. Contact him at xinyangli@tsinghua.edu.cn.

Mai Xu (Senior Member, IEEE) received the B.S. degree from Beihang University, Beijing, China, in 2003, the M.S. degree from Tsinghua University, Beijing, in 2006, and the Ph.D. degree from Imperial College London, London, U.K., in 2010. From 2010 to 2012, he was a Research Fellow with the Department of Electrical Engineering, Tsinghua University. Since 2013, he has been with Beihang University, where he was an Associate Professor and promoted to Full Professor in 2019. From 2014 to 2015, he was a Visiting Researcher of MSRA. He has authored or coauthored more than 200 technical papers in international journals and conference proceedings, such as International Journal of Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Journal of Selected Topics in Signal Processing, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, and AAAI. His main research interests include image processing and computer vision. He was the recipient of best/top paper awards of IEEE/ACM conferences, such as ACM MM. He was an Associate Editor for IEEE Transactions on Image Processing and IEEE Transactions on Multimedia, the Lead Guest Editor of IEEE Journal of Selected Topics in Signal Processing, and the Area Chair or a TPC Member of many conferences, such as ICME and AAAI. He is an elected Member of Multimedia Signal Processing Technical Committee and IEEE Signal Processing Society. Contact him at MaiXu@buaa.edu.cn.