Spatiotemporal Blind-Spot Network with Calibrated Flow Alignment for Self-Supervised Video Denoising

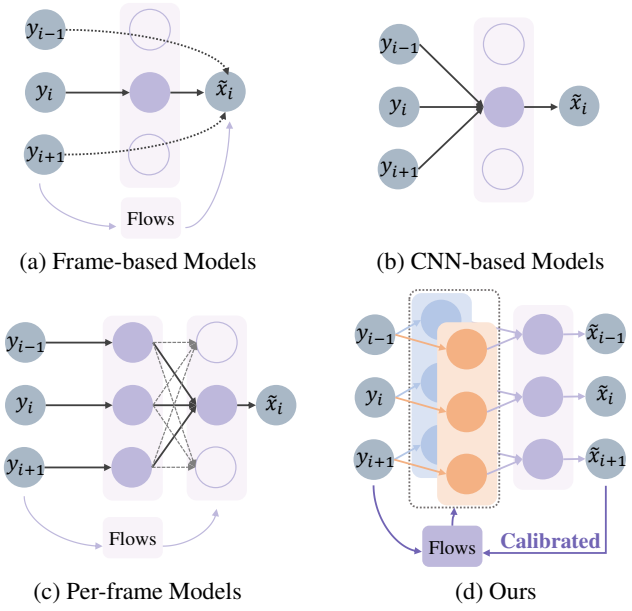
Zikang Chen, Tao Jiang, Xiaowan Hu, Wang Zhang, Huaqiu Li, Haoqian Wang*

Shenzhen International Graduate School, Tsinghua University
{czk23,jiang-t23,hu-xw19,zhangwan23,lihq23}@mails.tsinghua.edu.cn, wanghaoqian@tsinghua.edu.cn

Abstract

Self-supervised video denoising aims to remove noise from videos without relying on ground truth data, leveraging the video itself to recover clean frames. Existing methods often rely on simplistic feature stacking or apply optical flow without thorough analysis. This results in suboptimal utilization of both inter-frame and intra-frame information, and it also neglects the potential of optical flow alignment under self-supervised conditions, leading to biased and insufficient denoising outcomes. To this end, we first explore the practicality of optical flow in the self-supervised setting and introduce a **S**patio**T**emporal **B**lind-spot **N**etwork (**STBN**) for global frame feature utilization. In the temporal domain, we utilize bidirectional blind-spot feature propagation through the proposed blind-spot alignment block to ensure accurate temporal alignment and effectively capture long-range dependencies. In the spatial domain, we introduce the spatial receptive field expansion module, which enhances the receptive field and improves global perception capabilities. Additionally, to reduce the sensitivity of optical flow estimation to noise, we propose an unsupervised optical flow distillation mechanism that refines fine-grained inter-frame interactions during optical flow alignment. Our method demonstrates superior performance across both synthetic and real-world video denoising datasets.

Code — https://github.com/ZKCCZ/STBN



(a) Frame-based Models (b) CNN-based Models

(c) Per-frame Models (d) Ours

Figure 1: Illustrative comparison of frame sequence utilization strategies in self-supervised video denoising methods.

Introduction

Images captured under challenging environmental conditions, such as low lighting and slow shutter speeds, are often susceptible to various forms of noise and corruption. This issue is exacerbated in videos due to the typically higher shutter speeds, which not only degrades the overall quality of the video but also adversely affects subsequent computer vision tasks (Shen et al. 2020; Deng et al. 2022).

Given its critical role in computer vision, video denoising has witnessed significant advancements, largely driven by the application of deep learning techniques. Supervised video denoising methods, including Convolutional Neural Networks (CNNs) (Tassano, Delon, and Veit 2019, 2020), Recurrent Neural Networks (RNNs) (Chan et al. 2021; Li et al. 2022), and Transformer-based models (Liang et al. 2022, 2024), have made significant advancements. However, supervised video denoising methods rely heavily on labeled data, which is difficult and time-consuming to obtain. For example, obtaining the ground truth data of microscope videos and dynamic scenes is often impractical. This limitation restricts the applicability of supervised approaches in these contexts. Therefore, self-supervised methods have gained increasing attention as they eliminate the need for labeled training data. Grounded in the Noise2Noise assumption (Lehtinen et al. 2018), frame-based approaches (Ehret et al. 2019; Dewil et al. 2021) warp consecutive frames to create noise pairs for self-supervised training, as illustrated in Figure 1a. These methods heavily rely on precise optical flow estimation, which becomes particularly challenging in high-noise scenarios. The dependency can lead to severe artifacts in the warped images and an inefficient utilization of inter-frame redundancy. Additionally, CNN-based models (Sheth et al. 2021), depicted in Figure 1b,

*Corresponding author.

stack adjacent frames and employ blind-spot networks for self-supervised training. Per-frame models, as shown in Figure 1c, attempt to leverage all other aligned frames for each frame. However, this results in a computational complexity of $O(T^2)$. One possible approach is to align only a few adjacent frames (Zheng, Pang, and Ji 2023), yet this still compromises long-term information. These models are limited by their frame window size, restricting their ability to capture global temporal information.

Apart from the limited receptive field in both spatial and temporal domains, another significant issue lies in the efficiency and accuracy of optical flow utilization. The aforementioned methods that rely on frame-by-frame optical flow matching encounter a high computational complexity. Methods like RDRF (Wang et al. 2023) tackle these challenges by recurrently leveraging optical flow to capture long-term dependencies. However, as noted in their approach, their model is prone to overfitting, especially when dealing with real-world noisy data. Moreover, the reliance on unverified optical flow can introduce potential biases and errors, which need to be carefully examined within a self-supervised framework. Additionally, current methods are restricted to only access corrupted input video sequences for optical flow estimation, leading to suboptimal results due to the noise sensitivity of optical flow estimation.

To address the aforementioned challenges, we introduce a Spatiotemporal Blind-spot Network (STBN) to robustly handle both synthetic and real-world noise, as shown in Figure 1d. Our approach leverages inter-frame information through bidirectional alignment and propagation with the Blind-Spot Alignment (BSA) block for global temporal awareness. To integrate aligned temporal information and intra-frame features, we propose the Spatial Receptive Field Expansion (SRFE) module, which significantly enlarges the receptive field and further utilizes bidirectional spatial information. In the self-supervised setting, we discuss and calibrate feature alignment methods to ensure the consistency of noise distribution and independence, preserving the integrity of our self-supervised assumptions and avoiding potential biases. Moreover, considering the sensitivity of optical flow estimation to noise, we perform optical flow refinement using initially restored frames as pseudo-ground truth for knowledge distillation, enhancing noise robustness and improving spatiotemporal feature alignment and utilization. We summarize our contributions as follows:

- We propose a Spatiotemporal Blind-spot Network that effectively leverages inter-frame and intra-frame information through blind-spot temporal propagation and spatial fusion for self-supervised denoising for both synthetic and real noise.
- To ensure accurate utilization of temporal information in our self-supervised framework, we calibrate the multi-frame alignment paradigm to maintain global consistency of noise priors to prevent bias during training.
- The proposed knowledge distillation strategy in an unsupervised setting mitigates the sensitivity of optical flow to noise, thereby enhancing the precision of spatiotemporal feature utilization.
- Experimental results show that our method surpasses existing state-of-the-art self-supervised methods on various synthetic and real video noise datasets, demonstrating its superiority in video denoising tasks.

Related Work

Supervised Video Denoising

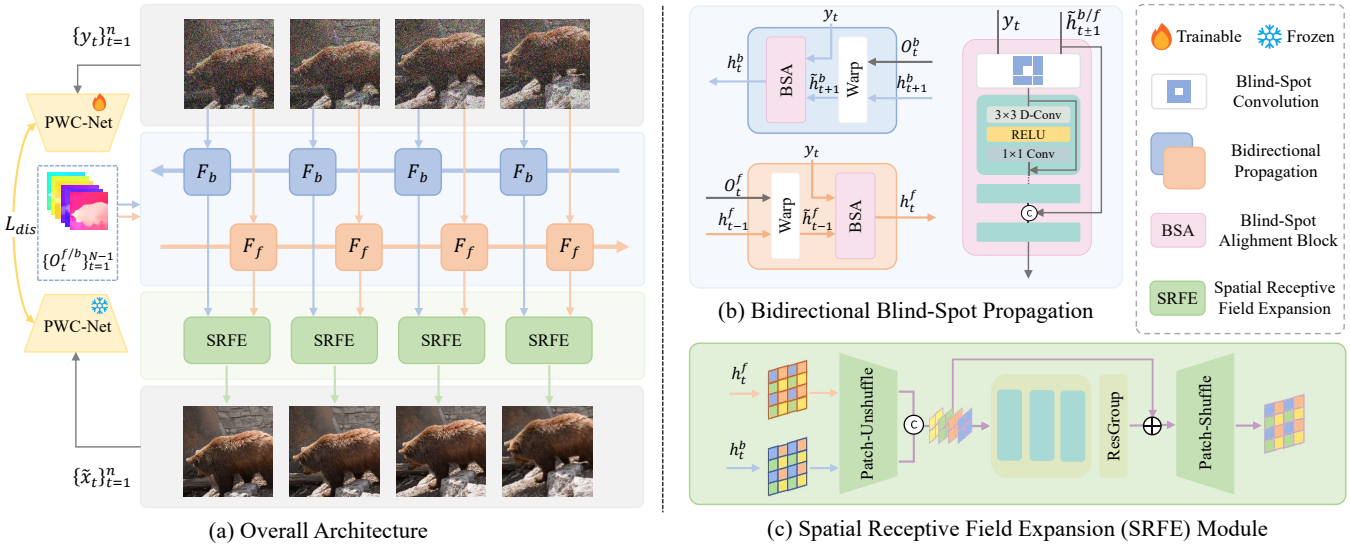
To leverage temporal redundancy to exploit inter-frame information, methods such as PaCNet (Ko, Lee, and Kim 2018) and VNLNet (Davy et al. 2019) utilize block matching combined with CNNs based on spatiotemporal neighborhoods, leading to high computational complexity. Alternatively, sliding window approaches like FastDVDnet (Tassano, Delon, and Veit 2020), an extension of DVDnet (Tassano, Delon, and Veit 2019), enhance efficiency by processing fixed-size consecutive frames through a two-level U-Net. Some methods incorporate optical flow for motion compensation, such as FloRNN (Li et al. 2022), which extends BasicVSR (Chan et al. 2021) by integrating future frame alignment for online denoising. VRT (Liang et al. 2024) processes video sequences in 2-frame clips with attention modules and optical flow for cross-clip interactions. RVRT (Liang et al. 2022) further enhances this by processing frames in parallel within a global recurrent framework.

Unsupervised Video Denoising

Traditional methods, such as VBM4D (Maggioni et al. 2012) based on BM3D (Dabov et al. 2007), use video filtering algorithms to find similar blocks for denoising. Recent deep learning-based approaches can be broadly categorized into noise-paired methods and blind-spot network methods. Frame2Frame (F2F) (Ehret et al. 2019) and Multi-Frame2Frame (MF2F) (Dewil et al. 2021), based on the Noise2Noise (N2N) (Lehtinen et al. 2018) assumption, align consecutive frames as noise pairs for denoising. ER2R (Zheng, Pang, and Ji 2023) extends the R2R (Pang et al. 2021) assumption, training by creating noise pairs through adding and subtracting noise from the original noisy videos when the specific noise distribution is known. It aligns each frame with others using a sliding window to reduce complexity, which leads to a significant loss of temporal information. Another approach extends blind-spot networks (Krull, Buchholz, and Jug 2019; Laine et al. 2019) to the video denoising domain. UDVD (Sheth et al. 2021) directly stacks a fixed length of adjacent frames into a blind-spot CNN. Although this method implicitly achieves feature alignment through a two-stage U-Net, it restricts the ability to utilize long-term temporal patterns by considering only frames within a limited window size. RDRF (Wang et al. 2023) employs 3D networks and a recurrent network based on blind spatial modulation to integrate features from near and far. However, this method is prone to overfitting, especially when dealing with raw video data.

Frame Alignment in Video Restoration

In video restoration, aligning highly correlated but temporally unsynchronized frames is crucial (Nah, Son, and Lee



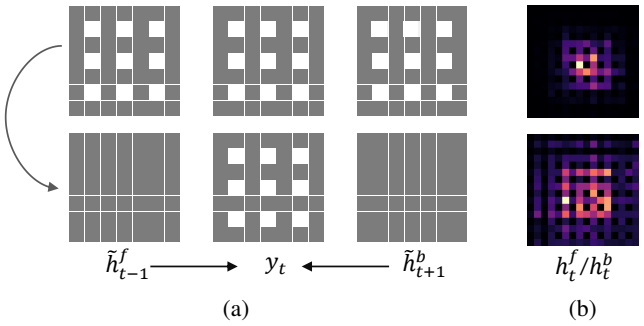
(a) Overall Architecture

(b) Bidirectional Blind-Spot Propagation

(c) Spatial Receptive Field Expansion (SRFE) Module

Figure 2: Illustration of the proposed method: (a) Overall architecture of STBN, including spatiotemporal feature aggregation and optical flow refinement. (b) The Bidirectional Blind-Spot Propagation utilizes the BSA block for global temporal awareness in both forward and backward propagation. (c) Detailed process of the Spatial Receptive Field Expansion module, which sequentially incorporates patch-shuffle, residual blocks, and patch-unshuffle to effectively enhance the spatial receptive field.

2019; Chan et al. 2021). Many methods use optical flow for frame alignment. BasicVSR (Chan et al. 2021) employs optical flow for recurrent feature propagation, and BasicVSR++ (Chan et al. 2022) uses it to guide offset learning. Task-specific optical flow is fine-tuned using models like SpyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018) for specific restoration tasks (Xue et al. 2019). Despite its efficiency in video restoration (Chan et al. 2021, 2022), the use of optical flow in self-supervised denoising has been less explored. UDVD (Sheth et al. 2021) achieves implicit alignment with a two-stage U-Net, while some methods (Yu et al. 2020) use trainable estimators for improved alignment. The applicability and effectiveness of optical flow alignment in self-supervised denoising remain to be explored.



(a) (b)

Figure 3: Visualization of (a) BSA block for temporal processing and (b) SRFE for spatial receptive field expansion.

Methodology

Let $\boldsymbol{y} \in \mathbb{R}^{T\times H\times W\times C}$ represent the noisy input frame sequence and $\boldsymbol{x} \in \mathbb{R}^{T\times H\times W\times C}$ denote the potentially clean target frame sequence, where T, H, W, and C are the video length, height, width, and channel, respectively. The overall framework of STBN is illustrated in Figure 2a. Initially, optical flow is predicted from the noisy video sequence and fed into the bidirectional blind-spot propagation module, where features are aligned within the Blind-Spot Alignment (BSA) block. The temporal information is then passed to the Spatial Receptive Field Expansion (SRFE) module, significantly expanding the receptive field of the blind-spots. The fused features are used to generate the final output and serve as pseudo-ground truth for further optical flow refinement.

Calibration of Frame Alignment

To achieve global temporal feature utilization, we employ optical flow for bidirectional feature warping. In this section, we examine the applicability of optical flow alignment methods within the self-supervised learning framework.

First, we propose the blind-spot network assumption for video sequences, where noise is pixel-independent both temporally and spatially, and pixel information can be inferred from the spatiotemporal context in the video. We assume that at the t-th frame $\boldsymbol{y}_t$, the receptive field for the i-th pixel $\boldsymbol{y}_{(t,i)}$, which acts as the blind-spot in our model, is denoted as $\boldsymbol{y}_{t,RF(i)}$. We define our model as the function as follows:

$$f\left(\boldsymbol{y}_{t,RF(i)}, warp(\boldsymbol{y}_k, \boldsymbol{O}_k); \boldsymbol{\theta}\right) = \boldsymbol{y}_{(t,i)}, \quad k \in \{1, 2, ..., T\} \setminus \{t\}, \quad (1)$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters we aim to train, $\boldsymbol{O}_k$ represents the estimated optical flow between two frames, and $warp$ represents the alignment operation. The

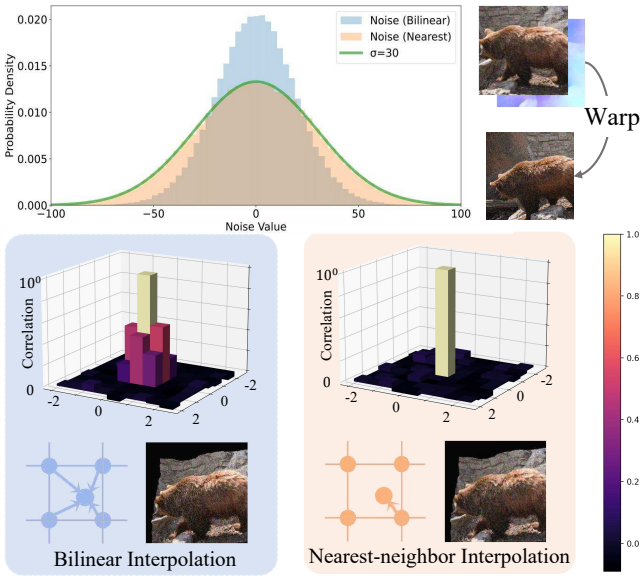


Figure 4: Visualization of noise distribution and correlation for two interpolation methods. Bilinear interpolation introduces spatial correlation and distorts the noise distribution, while nearest-neighbor interpolation preserves it.

model is trained by minimizing the empirical risk below:

$$\arg\min_{\theta} \sum_{t,i} L\left(f\left(\boldsymbol{y}_{(t,RF(i))}, warp(\boldsymbol{y}_k, \boldsymbol{O}_k); \boldsymbol{\theta}\right), \boldsymbol{y}_{(t,i)}\right). \tag{2}$$

The above formulation can be considered equivalent to the supervised training process. The detailed proof is provided in the supplementary material.

As shown in the above derivation, the inputs to f necessitate that both $\boldsymbol{y}_t$ and $warp(\boldsymbol{y}_k, \boldsymbol{O}_k)$, i.e., the noise from the current frame and the aligned frames, must remain pixel-independent both temporally and spatially. In optical flow alignment, bilinear and nearest-neighbor interpolation are two commonly employed methods. We use these as examples to illustrate the impact of alignment on noise characteristics and correlation. As shown in Figure 4, we performed forward warping on frames using these two methods, respectively. The same operation is applied on the ground truth data to calculate the noise distribution after interpolation. It can be observed that bilinear interpolation not only disrupts the distribution of noise but also introduces spatial correlations. This occurs because bilinear interpolation uses surrounding pixel information, performing a filtering-like operation on the image, which violates our self-supervised assumptions and leads to method failure. In contrast, nearest-neighbor interpolation preserves the original pixel values, maintaining the noise distribution and its independence. This is further demonstrated in our experiments.

Spatiotemporal Blind-Spot Feature Aggregation

To better utilize video frame sequences in both spatial and temporal domains, we design two distinct modules: the temporal module, which performs bidirectional alignment and propagation of features, and the spatial module, which significantly expands the receptive field to more effectively leverage the aligned frames. Together, these modules enable the model to achieve global awareness and enhance spatiotemporal feature integration.

Bidirectional Blind-Spot Propagation. To perform temporal feature alignment and propagation, we design a feature propagation and alignment module using blind-spot convolutions and dilated convolutions, as illustrated in Figure 2b. The input $\boldsymbol{y}_t$ from the t-th frame, along with the bidirectionally propagated features $\boldsymbol{h}^f_{t-1}$ or $\boldsymbol{h}^b_{t+1}$, which are warped to the current frame using optical flow, are then fed into the Blind-Spot Alignment (BSA) block for motion compensation. The entire process is as follows:

$$\begin{aligned} \boldsymbol{h}^f_t &= F_f\left(\boldsymbol{y}_t, warp(\boldsymbol{h}^f_{t-1}, \boldsymbol{O}^f_t)\right), \\ \boldsymbol{h}^b_t &= F_b\left(\boldsymbol{y}_t, warp(\boldsymbol{h}^b_{t+1}, \boldsymbol{O}^b_t)\right), \end{aligned} \tag{3}$$

where F_f, F_b denote forward and backward propagation, $\boldsymbol{O}^f_t, \boldsymbol{O}^b_t$ represent the bidirectional estimated optical flow.

During the alignment process, the BSA block is designed to maximally leverage the features from both forward and backward propagation. First, $\boldsymbol{y}_t$ and $\boldsymbol{h}$ are concatenated and then passed through a blind-spot convolution. The output is subsequently processed by modules that consist of a dilated convolution, an activation layer, and a 1×1 convolution. Although the features are well-aligned at this stage, they are not fully utilized. Therefore, we further concatenate the output with feature $\boldsymbol{h}$ and pass it through the blind-spot convolution block. Figure 3a shows the dependency between input and output pixels, with white pixels indicating regions independent of the central pixel and gray pixels representing convolution weights. This demonstrates that the BSA block effectively utilizes all temporal redundancy.

Spatial Receptive Field Expansion. Once the bidirectional features are aligned to $\boldsymbol{y}_t$, they inherently capture the temporal features of the entire sequence. To further leverage the aligned frames, we expand the receptive field under the blind-spot framework to utilize spatial domain information for enhanced image recovery. Inspired by (Jang et al. 2024; Li, Zhang, and Zuo 2024), we propose the Spatial Receptive Field Expansion (SRFE) Module, as illustrated in Figure 2c.

In the SRFE module, the forward features $\boldsymbol{h}^f$ and backward features $\boldsymbol{h}^b$ are first processed through a patch-unshuffle operation, and then stacked together to pass through several residual blocks, which ensure thorough feature fusion and enhance the model's ability to capture contextual information. Finally, the features are restored to their original size through a patch-shuffle operation, producing the output. The process can be represented as follows:

$$\tilde{\boldsymbol{x}}_t = \mathbf{SRFE}(\boldsymbol{h}^b_t, \boldsymbol{h}^f_t),\ t = 1, 2, ...T \tag{4}$$

As shown in Figure 3b, our strategy leads to a substantial increase in the receptive field, effectively integrating spatial information. This expansion enhances the model's capability to capture and utilize detailed spatial features.

Figure 5: Visual comparisons of different methods on synthetic noise data.

Dataset	σ	Traditional	Supervised			Unsupervised			
		VBM4D	DVDnet	FastDVDnet	FloRNN	UDVD	RDRF	ER2R$_s$	**STBN (Ours)**
Set8	10	36.05/-	36.08/0.9510	36.44/0.9540	37.57/0.9639	36.36/0.9510	36.67/0.9547	37.55/-	**37.24/0.9594**
	20	32.19/-	33.49/0.9182	33.43/0.9196	34.67/0.9379	33.53/0.9167	34.00/0.9251	34.34/-	**34.41/0.9322**
	30	30.00/-	31.68/0.8862	31.68/0.8889	32.97/0.9138	31.88/0.8865	32.39/0.8978	32.45/-	**32.76/0.9072**
	40	28.48/-	30.46/0.8564	30.46/0.8608	31.75/0.8911	30.72/0.8595	31.23/0.8725	31.09/-	**31.57/0.8837**
	50	27.33/-	29.53/0.8289	29.53/0.8351	30.80/0.8696	29.81/0.8349	30.31/0.8490	30.05/-	**30.62/0.8608**
	avg	30.81/-	32.29/0.8881	32.31/0.8917	33.55/0.9153	32.46/0.8897	32.92/0.8998	33.10/-	**33.32/0.9087**
DAVIS	10	37.58/-	38.13/0.9657	38.71/0.9672	40.16/0.9755	39.17/0.9700	39.54/0.9717	39.52/-	**40.35/0.9613**
	20	33.88/-	35.70/0.9422	35.77/0.9405	37.52/0.9564	35.94/0.9428	36.40/0.9473	36.49/-	**37.67/0.9606**
	30	31.65/-	34.08/0.9188	34.04/0.9167	35.89/0.9440	34.09/0.9178	34.55/0.9245	34.60/-	**36.00/0.9454**
	40	30.05/-	32.86/0.8962	32.82/0.8949	34.66/0.9286	32.79/0.8949	33.23/0.9032	33.29/-	**34.73/0.9296**
	50	28.80/-	31.85/0.8745	31.86/0.8747	33.67/0.9131	31.80/0.8739	32.20/0.8832	32.25/-	**33.70/0.9138**
	avg	32.39/-	34.52/0.9195	34.64/0.9188	36.38/0.9435	34.76/0.9199	35.18/0.9260	35.23/-	**36.49/0.9451**

Table 1: Quantitative comparison of PSNR/SSIM for Gaussian denoising on the Set8 and DAVIS datasets. The best results for unsupervised methods are in bold. Note that ER2R$_s$ utilizes the same video sequence for both training and testing.

Flow Refinement in Noise

Optical flow estimation, which is also sensitive to noise, significantly impacts the accuracy of temporal alignment. To address the issue of imprecise optical flow estimation caused by corrupted input images, we introduce a knowledge distillation approach that uses pseudo-ground truth for optical flow refinement.

We define the method of optical flow estimator as $\mathcal{E}(\cdot)$. Once the training achieves preliminary effectiveness, we generate clean video sequences $\tilde{x}$ using a frozen-parameter $\mathcal{E}_{\text{fix}}(\cdot)$ to serve as pseudo-ground truth. These sequences and the original video frame $\boldsymbol{y}$ are used for optical flow estimation respectively as follows:

$$\tilde{\boldsymbol{O}}_t^f = sg(\mathcal{E}_{\text{fix}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_{t+1})),\ \boldsymbol{O}_t^f = \mathcal{E}(\boldsymbol{y}_t, \boldsymbol{y}_{t+1}), \quad (5)$$

where $sg(\cdot)$ is the stop gradient operation. This accurate optical flow $\tilde{\boldsymbol{O}}_t^f$ treated as pseudo-ground truth is then to guide the refinement of the optical flow estimation in noisy video sequences. We optimize the original optical flow estimator using the following loss function:

$$\mathcal{L}_{dis} = \sum_t \left\| \tilde{\boldsymbol{O}}_t^f - \boldsymbol{O}_t^f \right\|_1. \quad (6)$$

The distillation loss, scaled by a small coefficient α as a constraint, is jointly trained with our model. This distillation approach enhances the performance of the optical flow estimator in the presence of noise, thereby improving overall temporal alignment and benefiting the entire model.

Experiments

Implementation Details

We conduct experiments on both synthetic and real raw noise. For synthetic noise, following (Sheth et al. 2021; Wang et al. 2023), we train our model with negative log-likelihood loss $\mathcal{L}_{log}$ and test them with posterior inference (Laine et al. 2019). For real raw noise, we use $\mathcal{L}_2$ loss for self-supervised training. For the optical flow estimator, we use the pre-trained PWC-Net (Sun et al. 2018) as our initial optical flow extractor. The distillation loss is introduced with $\alpha = 5 \times 10^{-4}$. Training sequences are spatially cropped to a size of 96×96 and temporally to a length of $T = 10$ for synthetic data and $T = 7$ for real data. All experiments are carried out using the Adam optimizer with an initial learning rate of 1×10^{-4} on a single RTX 3090 GPU. We used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as evaluation metrics.

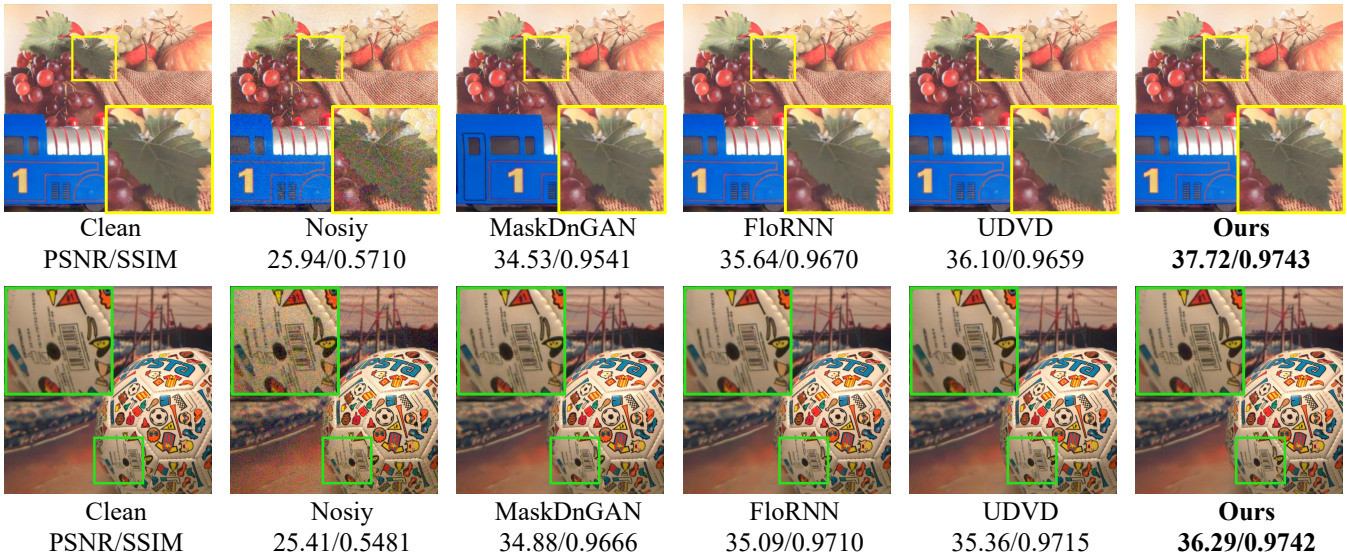


Figure 6: Visual comparisons on CRVD dataset. The results have been converted to the sRGB domain for visualization.

ISO	Supervised				Unsupervised			
	FastDVDnet	RViDeNet	MaskDnGAN	FloRNN	UDVD	RDRF	ER2R$_p$	**STBN (Ours)**
1600	43.43/0.9866	47.74/0.9938	47.52/0.9941	48.81/0.9956	48.02/0.9982	48.38/0.9983	49.14/-	**49.27/0.9988**
3200	42.91/0.9844	45.91/0.9911	45.88/0.9914	47.05/0.9933	46.44/0.9980	46.86/0.9981	47.51/-	**47.58/0.9985**
6400	40.29/0.9793	43.85/0.9880	44.14/0.9886	45.09/0.9910	44.74/0.9972	45.24/0.9975	45.61/-	**45.75/0.9980**
12800	36.05/0.9613	41.20/0.9819	41.48/0.9834	42.63/0.9866	42.21/0.9966	42.72/0.9969	43.03/-	**43.36/0.9976**
25600	36.50/0.9400	41.17/0.9821	40.79/0.9819	42.19/0.9872	42.13/0.9951	42.25/0.9948	42.91/-	**42.91/0.9972**
avg	39.84/0.9703	43.97/0.9874	43.96/0.9880	45.15/0.9907	44.71/0.9970	45.09/0.9971	45.64/-	**45.77/0.9980**

Table 2: Quantitative comparison of PSNR/SSIM on the CRVD dataset. The best results for unsupervised methods are in bold. Note that ER2R$_p$ utilizes extra noise distribution priors to generate noise pairs during the training process.

Experiments on Synthetic Noise

In our experiments on synthetic noise, we utilize DAVIS dataset (Pont-Tuset et al. 2017) and Set8 (Tassano, Delon, and Veit 2019) dataset. To generate noisy video sequences, additive white Gaussian noise (AWGN) with a standard deviation $\sigma \in [5, 55]$ is introduced to the training dataset. We compare our method with a range of benchmarks, including the non-learning method VBM4D (Maggioni et al. 2012), supervised approaches such as FastDVDnet (Tassano, Delon, and Veit 2020), PaCNet (Vaksman, Elad, and Milanfar 2021), and FloRNN (Li et al. 2022), as well as unsupervised methods like UDVD (Sheth et al. 2021), RDRF (Wang et al. 2023), and ER2R (Zheng, Pang, and Ji 2023).

Quantitative Comparison. Table 1 reports the PSNR and SSIM of different methods on the DAVIS testing set and Set8 datasets under different noise levels. Note that ER2R utilizes the same video sequence for both training and testing. Our model outperforms RDRF by an average PSNR of 1.31 dB and 0.4 dB on two different datasets and is highly comparable to the supervised method FloRNN. The results demonstrate the effectiveness of our global spatiotemporal perception and refined optical flow alignment, highlighting the advantages of our self-supervised method. Figure 5 illustrates our qualitative results, showing that our method restores corrupted text more accurately compared to existing approaches, which demonstrates the effectiveness of our approach in preserving fine details.

Experiments on Real Raw Noise

We evaluate our method using the CRVD dataset (Yue et al. 2020), a real-world video denoising dataset captured in the raw domain, to assess our performance on real-world noise. This dataset comprises 6 indoor scenes for training and 5 indoor scenes for testing, with each scene consisting of 7 frames with 10 different noise realizations captured at five different ISO levels. We compare our method against several approaches, including supervised methods FastDVDnet (Tassano, Delon, and Veit 2020), RViDeNet (Yue et al. 2020), MaskDnGAN (Paliwal, Zeng, and Kalantari 2021), and FloRNN (Li et al. 2022), as well as unsupervised methods such as UDVD (Sheth et al. 2021), RDRF (Wang et al. 2023), and ER2R (Zheng, Pang, and Ji 2023). For a fair comparison, both training and testing are performed on the test

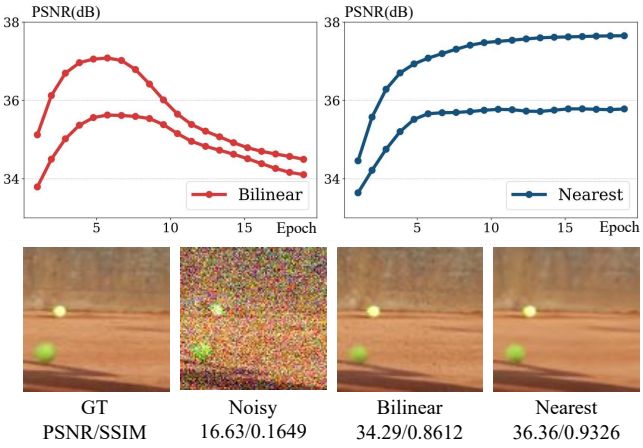


Figure 7: Visualizations of experimental results during training with different warping methods.

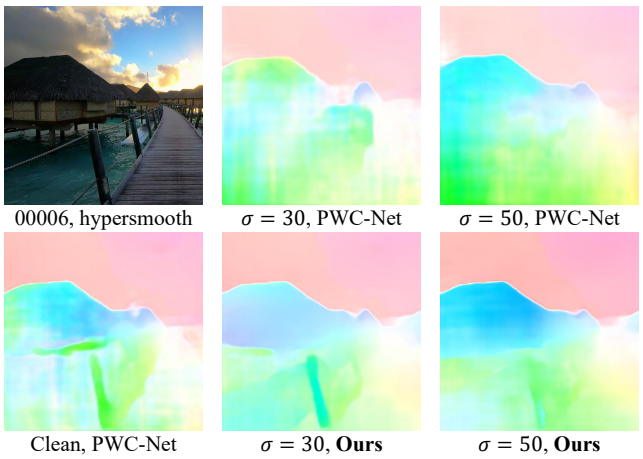


Figure 8: Visualization of optical flow for the initial estimator compared to our refined results.

sequences as employed in previous unsupervised methods.

Quantitative Comparison. Table 2 reports our results on the CRVD dataset. Note that ER2R utilizes prior noise information by creating noise pairs during training, whereas ours rely solely on noisy images. Our model demonstrates superior performance, surpassing RDRF by 0.68 dB under identical settings. While RDRF requires meticulous tuning to prevent overfitting with limited samples, our method leverages well-calibrated optical flow alignment and a robust spatiotemporal blind-spot network, which enables precise global information aggregation to improve denoising results. Our approach exceeds ER2R by an average of 0.13 dB even though we have very limited data. Under the above training setting, we outperform the supervised method FloRNN by an average of 0.62 dB. Given the challenges of obtaining ground truth in real noise scenarios, our unsupervised approach demonstrates greater practical utility. As illustrated in Figure 6, our method better preserves high-frequency details that others often lose during the denoising process.

Component	Methods			
Propagation	✓	✓	✓	✓
BSA Block		✓	✓	✓
SRFE Module			✓	✓
Optical Refinement				✓
PSNR	32.14	32.49	32.68	**32.76**
SSIM	0.8942	0.9037	0.9068	**0.9072**

Table 3: Ablation study of model components.

Analysis of the Proposed Method

Ablation study. We perform ablation studies on the Set8 dataset with Gaussian noise level σ=30, as detailed in Table 3. Starting with temporal feature propagation alone, incorporating the BSA block enhances temporal feature utilization, improving PSNR by 0.35 dB. Adding the SRFE module further leverages spatial information, resulting in an additional 0.19 dB increase in PSNR. Finally, introducing optical flow refinement provides a further improvement of 0.08 dB in PSNR. These results demonstrate that gradually utilizing temporal and spatial features and refining alignment incrementally enhances model performance.

Feature Alignment Strategy. Figure 7 presents the results of experiments conducted on two samples from the Set8 dataset using bilinear interpolation and nearest-neighbor interpolation. The former led to a decrease in PSNR during training, which aligns with our conclusions that bilinear interpolation disrupts the noise structure, thereby violating our blind-spot assumption. Consequently, bilinear interpolation produced poor visual results.

Optical Flow Refinement. We visualize the refined optical flow produced by our proposed method for noise levels $\sigma = 30$ and $\sigma = 50$ as shown in Figure 8. The optical flow estimator benefits from knowledge distillation guided by generated pseudo-ground truths in the training process, leading to more accurate optical flow predictions under noisy conditions. Consequently, our alignment module achieves improved matching accuracy, which in turn contributes to the superior performance of our denoising model.

Conclusion

In this paper, we introduce STBN for self-supervised video denoising. We validate and calibrate the multi-frame alignment paradigm within a self-supervised framework to ensure the global consistency of the noise prior, thereby mitigating training bias. Our proposed spatiotemporal blind-spot feature aggregation preserves long-range temporal dependencies and enhances spatial receptive fields for comprehensive global perception. Additionally, our unsupervised optical flow refinement reduces sensitivity to noise, improving the precision of spatiotemporal feature utilization. Experimental results demonstrate that our method surpasses existing unsupervised approaches and shows strong comparability to supervised methods, demonstrating great potential.

Acknowledgments

This work is supported by the Shenzhen Science and Technology Project under Grant (JCYJ20220818101001004).

References

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. 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*, 4947–4956.

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

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. In *2007 IEEE international conference on image processing*, volume 1, I–313. IEEE.

Davy, A.; Ehret, T.; Morel, J.-M.; Arias, P.; and Facciolo, G. 2019. A non-local CNN for video denoising. In *2019 IEEE international conference on image processing (ICIP)*, 2409–2413. IEEE.

Deng, X.; Wang, P.; Lian, X.; and Newsam, S. 2022. NightLab: A dual-level architecture with hardness detection for segmentation at night. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16938–16948.

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

Ehret, T.; Davy, A.; Morel, J.-M.; Facciolo, G.; and Arias, P. 2019. Model-blind video denoising via frame-to-frame training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11369–11378.

Jang, H.; Park, J.; Jung, D.; Lew, J.; Bae, H.; and Yoon, S. 2024. PUCA: patch-unshuffle and channel attention for enhanced self-supervised image denoising. *Advances in Neural Information Processing Systems*, 36.

Ko, K.; Lee, J.-T.; and Kim, C.-S. 2018. PAC-Net: pairwise aesthetic comparison network for image aesthetic assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2491–2495. IEEE.

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

Laine, S.; Karras, T.; Lehtinen, J.; and Aila, T. 2019. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

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

Li, J.; Zhang, Z.; and Zuo, W. 2024. TBSN: Transformer-Based Blind-Spot Network for Self-Supervised Image Denoising. *arXiv preprint arXiv:2404.07846*.

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

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

Maggioni, M.; Boracchi, G.; Foi, A.; and Egiazarian, K. 2012. Video denoising, deblocking, and enhancement through separable 4-D nonlocal spatiotemporal transforms. *IEEE Transactions on image processing*, 21(9): 3952–3966.

Nah, S.; Son, S.; and Lee, K. M. 2019. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8102–8111.

Paliwal, A.; Zeng, L.; and Kalantari, N. K. 2021. Multi-stage raw video denoising with adversarial loss and gradient mask. In *2021 IEEE International Conference on Computational Photography (ICCP)*, 1–10. IEEE.

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

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.

Shen, Y.; Ji, R.; Chen, Z.; Hong, X.; Zheng, F.; Liu, J.; Xu, M.; and Tian, Q. 2020. Noise-aware fully webly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11326–11335.

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

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

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

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

Vaksman, G.; Elad, M.; and Milanfar, P. 2021. Patch craft: Video denoising by deep modeling and patch matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2157–2166.

Wang, Z.; Zhang, Y.; Zhang, D.; and Fu, Y. 2023. Recurrent self-supervised video denoising with denser receptive field. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7363–7372.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127: 1106–1125.

Yu, S.; Park, B.; Park, J.; and Jeong, J. 2020. Joint learning of blind video denoising and optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 500–501.

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

Zheng, H.; Pang, T.; and Ji, H. 2023. Unsupervised deep video denoising with untrained network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3651–3659.