
BFCAI
Faculty of Computers and Artificial Intelligence

# A Survey on Real-Time Image/Video Denoising and Deblurring

R. Nabil, A. ELsayed, S. Ahmed, A. Mohamed and O. Abdelhameed, "real-time image/video denoising and deblurring," supervised by Dr.M.kamal, Faculty of Computers & Artificial Intelligence (BFCAI), Egypt.

## Abstract

Real-time image and video denoising and deblurring are challenging due to computational complexity and diverse distortions. Recent deep learning advancements use spatial-temporal modeling, attention mechanisms, and generative models (e.g., Residual Dense RNNs, Swin Transformers) to enhance restoration and enable real-time performance [1, 2]. Techniques like global-local attention and motion vectors improve adaptability [1], while new datasets address real-world benchmarks [6]. Joint frameworks tackle deblurring, denoising, and low-light enhancement [1, 3], with lightweight models (e.g., PTFN, RCD) supporting mobile deployment [6].

## 1.Introduction

In recent years, the demand for fast and efficient image and video processing techniques has increased, especially with the widespread use of cameras in mobile devices, surveillance systems, and digital media. These media face major challenges such as denoising and motion blur removal (deblurring), which directly affect the quality and effectiveness of the content in applications like computer vision and augmented reality [1], [6].

With advancements in artificial intelligence, solutions based on deep neural networks, transformers, and diffusion models have emerged to enhance the quality of images and videos in real-time. For example, models have been developed that combine multiple tasks such as multi-frame interpolation with simultaneous motion blur removal [1].

Hybrid transformer techniques have also contributed to improving the recovery of sharp video details, thereby enhancing the accuracy of motion blur removal [2]. Additionally, diffusion models have

proven effective in addressing image and video denoising, as demonstrated in studies like Swin-Diff [3] and Denoising Diffusion Models for Plug-and-Play Image Restoration [4].

These studies aim to provide integrated solutions that operate efficiently in real-time, with control over the quality of restoration and processing speed, as developed in the paper Real-time Controllable Denoising for Image and Video [5], along with improvements in temporal information fusion to enhance video denoising quality, as seen in the paper Towards High-Quality Real-Time Video Denoising with Pseudo Temporal Fusion Network [6].

With the above improvement, the proposed method can achieve better performance with less computational cost against SoTA deblurring methods, as illustrated in Fig. 1(a). Due to making full use of spatiotemporal dependency of video signal, our method is exceptionally good at restoring high-frequency details of the blurry frame compared with SoTA video deblurring methods, as shown in Fig. 1(b).

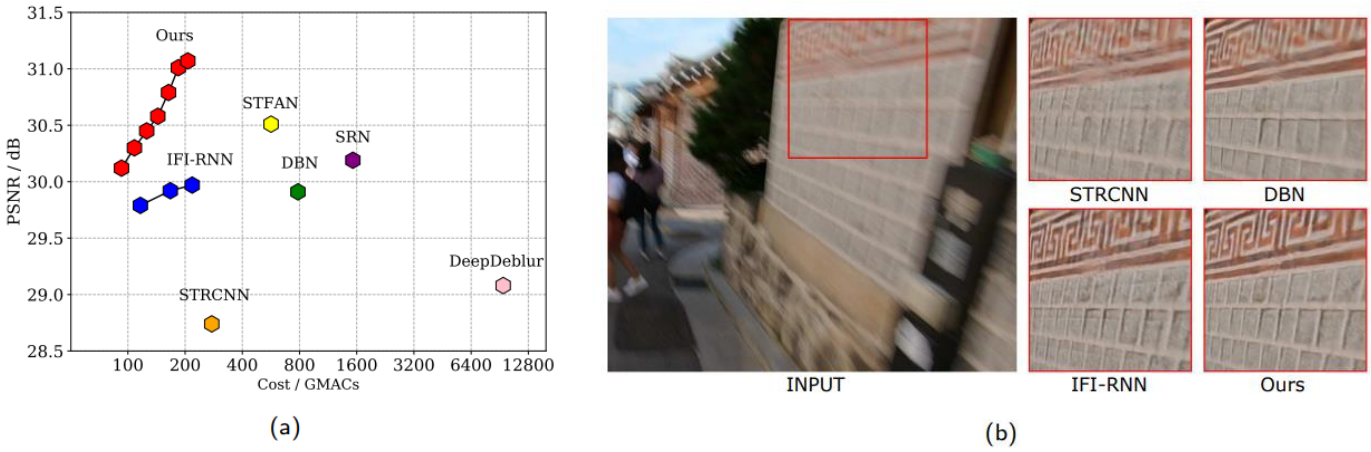



Fig. 1 A comparison of network efficiency on video deblurring.

# 2 .Background & Related work (Literature Review)

## 2.1. Denoising

Traditional image and video denoising methods have historically depended on prior assumptions such as sparse image priors, non-local similarity, and other related techniques. These approaches established the groundwork for early denoising efforts. The emergence of deep learning has shifted the paradigm, with learning-based methods achieving state-of-the-art results. Early attempts, such as those employing multi-layer perceptrons, demonstrated competitive performance compared to traditional methods like BM3D. Recent advancements have seen the dominance of convolutional neural network-based techniques and Transformer-based approaches in image and video denoising. However, these methods primarily focus on designing novel network architectures to enhance denoising performance, often producing a single output without the flexibility to adjust denoising levels based on user feedback. This limitation hinders their practical application in real-world scenarios. Additionally, while techniques like pruning and

quantization can accelerate these neural network-based methods, their computational heaviness restricts real-time denoising control.

## 2.2. Controllable Denoising

Conventional deep-learning methods for image and video denoising typically generate fixed results with a predetermined restoration level. Recent developments have introduced controllable denoising techniques, enabling users to adjust the restoration effect without retraining the network. Methods like DNI and AdaFM leverage the observation that filters learned by models trained at different restoration levels share similar visual patterns. DNI interpolates parameters between related networks to achieve smooth, continuous restoration effects, while AdaFM applies feature modulation filters after each convolutional layer. CFSNet proposes an adaptive learning strategy using interpolation coefficients to couple intermediate features between a main branch and a tuning branch. In contrast, alternative approaches treat modulation as a conditional image restoration problem, employing joint training strategies. CUGAN introduces a GAN-based framework to address the over-smoothing issue common in PSNR-oriented methods. However, these controllable methods are limited to training with synthetic degradations, requiring explicit degradation levels during training. When applied to real-world data, methods trained for blind Additive White Gaussian Noise often overfit, leading to significant performance drops. Moreover, these techniques rely on auxiliary conditional networks, necessitating a separate network inference for each target restoration level.

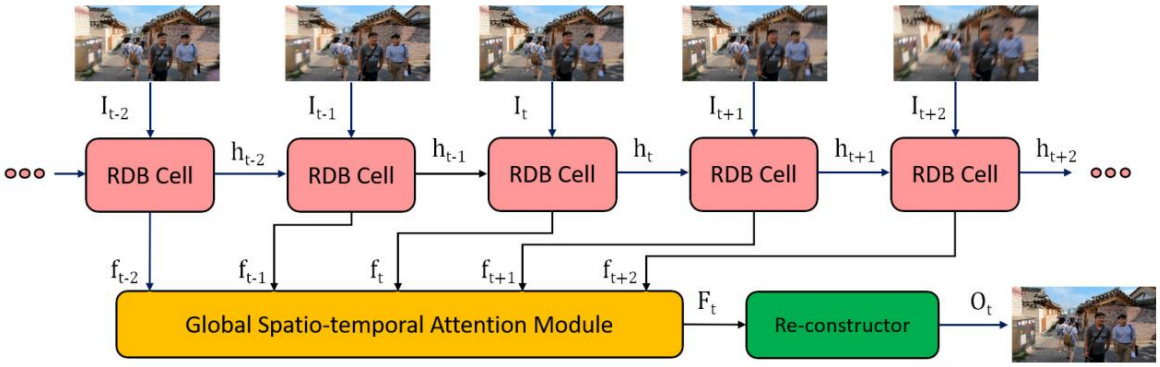
## 2.3. Image and Video Deblurring

Deep learning has significantly advanced single image deblurring, with methods being widely explored. One approach introduced a scale-recurrent network using a coarse-to-fine scheme to extract multi-scale features from blurry images, while another proposed a deep hierarchical multi-patch network inspired by spatial pyramid matching for handling blurry images effectively. Another utilized an asymmetric autoencoder and a fully-connected network for self-supervised image deblurring, and a reevaluation of the coarse-to-fine approach proposed a fast and accurate deblurring network.

For video deblurring, a spatial-temporal recurrent network with a dynamic temporal blending layer was developed to restore latent frames. Another method enhanced this by incorporating an optical flow estimation step to align and aggregate information across neighboring frames. A spatially variant RNN integrated with CNNs was employed to address spatially variant blur in dynamic scenes, while a combination of non-local blocks, recursive blocks, and a temporal loss function captured complex spatio-temporal patterns. Deformable convolution in a pyramid manner was introduced for better temporal alignment, and simultaneous estimation of optical flow and latent frames using a temporal sharpness prior fed estimated flow back into the reconstruction network. A spatiotemporal pyramid module captured multi-scale spatial and temporal information, and temporal-spatial and channel attention mechanisms modeled complex blur patterns. Enhanced RNN cells with residual dense blocks enabled efficient spatial feature extraction, adding a global spatio-temporal attention module to fuse hierarchical features from past and future frames.

Existing video deblurring methods often assume consecutive blurry frames, which may not align with real-world scenarios where some frames remain sharp. One approach addressed this by detecting sharp frames and using them to guide the restoration of blurry frames, though their simple concatenation approach limits the exploitation of sharp textures. This work aims to propose a new framework to better leverage sharp frames for improved video deblurring.

Fig. 2 Framework of the proposed efficient spatio-temporal recurrent neural network.




In this paper, we adopt a RNN framework similar to [1]. Our method is different from [1] in that we integrate RDB into the RNN cell in order to exploit the potential of the RNN cell through feature reusing and generating hierarchical features for the current frame. Furthermore, we propose a GSA module to selectively merge effective hierarchical features from both past and future frames, which enables our model to utilize the spatio-temporal information more efficiently.

# 3.Challenges

The field of video deblurring and denoising faces several persistent challenges driven by the complexity of real-world scenarios and the shortcomings of current methods. Below is a summary of key challenges framed generally to reflect broader trends in recent research.

## 3.1. Dynamic and Unknown Motion Blur

Challenge: Motion blur typically found in real-world videos results from uncontrolled camera exposure time and fast-moving objects. Due to a lack of prior knowledge about exposure time or motion paths, any attempt to model such blur becomes highly sophisticated.

Research Directions: Temporal dependencies in RNNs and hybrid transformers are being explored for capturing dependencies between frames. Multi-frame interpolation and alignment are becoming more popular for multi-frame retrieval.

## 3.2. Speed vs Quality

Challenge: The most modern models, like diffusion-based designs, achieve the highest quality restoration; however, their repetitive inference steps are costly in terms of computation, making real-time use very difficult.

Research Directions: Simplified architectures like pseudo-temporal fusion networks, along with model pruning, aim to decrease latency. Also, adaptive inference that deals with critical frames or regions first is being studied.

## 3.3. Multi-Task Learning's Competing Needs

Challenge: Solving a combination of deblurring, denoising, and low-light enhancement simultaneously usually creates conflicting goals. For instance, brightening an image can increase noise significantly, and aggressive denoising can remove fine details.

Research Directions: It is suggested that modular architectures with task-specific branching and a shared feature extraction backbone aim to resolve the balance between specialization and synergetic cooperation. Task-driven dynamic attention allows allocation of computational resources depending on the importance of a task.

## 3.4. Synthetic-to-Real Domain Gap

Challenge: Many datasets are synthesized and do not consider real-world scenarios. Thus, real-world videos consistently include noise, have erratic lighting, or diverse motion patterns.

Research Directions: Focus shifts onto self-supervised and unsupervised learning methods aimed towards unlabeled real-world data. In addition, models are trained to imitate the degradation of real-world data in diverse ways, including training with diffusion models.

## 3.5. Recovering Motion and Shape from Single Frames

Challenge: Challenge: Single-image deblurring is an ill-posed problem, attempting to reconstruct accurate motion trajectories or shapes for fast-moving objects that require inferring a plethora of information.

Research Directions: Increasingly popular are physics-informed models which combine deep learning with motion equations or optical flow priors. Probabilistic approaches, including diffusion models, propose estimating plausible trajectories by generating multiple hypotheses that need to be validated.

## 3.6. Integration of Domain Knowledge

Challenge: Effectively utilizing domain-specific a priori (e.g., video compression standards or spatiotemporal redundancies) comes with its own set of issues, particularly under-engineering the model.

Research Directions: Self-supervised processes that integrate custom layers built to replicate certain domain-specific processes, such as compression, are being explored alongside reinforcement learning.

## 3.7. Real-Time Controllability

Challenge: Applications are only interactive when full user control of the level of deblurring/denoising is available in real-time. Most models are not designed to incorporate style modifications by users, without significant sacrifices to acceleration.

Research Directions: The use of controllable output parametric modulation layers alongside multi-task designs enables real-time controls to respond as needed. Concentration on lightweight networks designed for edge devices is another area.

# 4.Methods

Zhong et al. (2021) propose the Efficient Spatio-Temporal Recurrent Neural Network (ESTRNN) for real-time video deblurring, designed to balance high-quality restoration with low computational demands, suitable for devices like smartphones. The methodology comprises three main components:

## 4.1.RDB-based RNN Cell

To efficiently extract spatial features, the authors integrate Residual Dense Blocks (RDBs) into the RNN cell. The cell processes the current blurry frame It and the previous hidden state ht−1. A downsampling operation concatenates these inputs to produce shallow feature maps:

$$f_t^D = \mathrm{CAT}(DS(I_t), h_{t-1}),$$

where CAT(·) denotes concatenation, and DS(·) is a downsampling operation using 5x5 convolutional layers and an RDB module. A series of RDB modules generates a feature set f R t = {f R1 t , . . . , f RN t }, where N is the number of RDB modules. Hierarchical features ft are obtained by fusing these features with a 1x1 convolutional layer:

$$f_t = \mathrm{Conv}(\mathrm{CAT}(f_t^R)).$$

The hidden state is updated as:

$$h_t = H(f_t),$$

where H is a function comprising a 3x3 convolutional layer and an RDB module, facilitating temporal information transfer.

## 4.2 Global Spatio-Temporal Attention (GSA) Module

The GSA module enhances deblurring by fusing hierarchical features from the current frame ft with those from two past and two future frames (ft−2, ft−1, ft+1, ft+2). Inspired by the Squeeze-and-Excitation block, it filters effective features from neighboring frames:

$$f_{t+i}^c = \mathrm{CAT}(f_t, f_{t+i}),$$

$$f_{t+i}^e = L(\mathrm{GAP}(f_{t+i}^c)) \otimes P(f_{t+i}^c),$$

where i ∈ {−2, −1, 1, 2}, GAP(·) is global average pooling, L(·) involves linear transformations with ReLU and Sigmoid activations, P(·) uses 1x1 convolutions, and ⊗ denotes element-wise multiplication. The fused output Ft is:

$$F_t = \mathrm{Conv}(\mathrm{CAT}(f_{t-2}^e, f_{t-1}^e, f_{t+1}^e, f_{t+2}^e, f_t)).$$

This output is upsampled by deconvolutional layers to produce the deblurred frame Ot.

### 4.3 Beam-Splitter Deblurring (BSD) Dataset

To address the scarcity of real-world datasets, the authors developed the BSD dataset using a beamsplitter system with two synchronized cameras. One camera captures blurry videos with longer exposure times, and the other captures sharp videos. A center-aligned exposure scheme and a 12.5% neutral density filter ensure alignment and photometric consistency. The dataset includes 60 training, 20 validation, and 20 test sequences at 640x480 resolution, with 100–150 frames per sequence, across three blur settings (1ms-8ms, 2ms-16ms, 3ms-24ms). Synthetic datasets are generated by averaging high-frame-rate frames:

$$I_b = \mathrm{CRF}\left(\frac{1}{N}\sum_{n=1}^{N} S^n\right),$$

where Ib is the blurry frame, S n is the n-th sharp frame, N is the number of frames, and CRF is the camera response function.

### 4.4 Training Configuration

The ESTRNN model was trained on GOPRO and REDS datasets using the ADAM optimizer (β1 = 0.9, β2 = 0.999), an initial learning rate of 10−4 , and MSE loss:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{TCHW}\sum_{t=1}^{T}\|O_t - O_t^{GT}\|_2^2,$$

where T, C, H, W are the number of frames, channels, height, and width, and OGT t is the ground truth. For BSD, a cosine annealing schedule (learning rate 3 × 10−4 ) and Charbonnier loss were used:

$$\mathcal{L}_{\mathrm{char}} = \frac{1}{TCHW}\sum_{t=1}^{T}\left\|\sqrt{(O_t - O_t^{GT})^2 + \epsilon^2}\right\|,$$

with ϵ = 1 × 10−3 , a mini-batch size of 8, and subsequence length of 8.

## 5 .Experiments and Results

The ESTRNN model was evaluated on synthetic (GOPRO, REDS) and real-world (BSD) datasets, outperforming methods like STRCNN, DBN, IFI-RNN, CDVD-TSP, and PVDNet:

1. Synthetic Datasets: On GOPRO, ESTRNN (B15C70) achieved a PSNR of 30.12 dB and SSIM of 0.8929, and on REDS, 31.94 dB and 0.8962, with computational costs of 92.57–125.55 GMACs, significantly lower than CDVD-TSP (5211.28 GMACs) and PVDNet (1754.90 GMACs).

2. BSD Dataset: On BSD, ESTRNN (B15C80) recorded PSNR/SSIM of 33.36/0.937 (1ms-8ms), 31.95/0.925 (2ms-16ms), and 31.39/0.926 (3ms-24ms), demonstrating robust performance across varying blur intensities. Visual results showed clearer restoration, especially in complex motion scenarios.

3. Cross-Validation: Models trained on BSD generalized well to synthetic datasets (e.g., PSNR/SSIM of 26.46/0.817 on GOPRO), while synthetic-trained models (e.g., GOPRO: 19.48/0.598 on BSD) produced artifacts on BSD. A 2000 fps synthetic dataset also underperformed, highlighting BSD's superior real-world applicability due to its complex noise and blur characteristics.

4. Real-World Testing: On DVD dataset videos and iPhone 13 footage, BSD-trained ESTRNN produced sharper results with fewer artifacts compared to synthetic-trained models, which struggled with dynamic regions.

# 6.Conclusion and Future

While techniques like Residual Dense RNNs and Swin Transformers improve performance, they often struggle with real-world data, such as irregular noise or sharp frames in videos. Research is moving toward unified frameworks that address multiple tasks, like denoising and deblurring, with lightweight models like PTFN and RCD for mobile deployment.

## Future Directions

It seems likely that future research will focus on improving performance on real-world data, developing new datasets, and reducing computational costs, potentially leading to more practical and flexible solutions.

## REFERENCES


[1] C. Shang *et al*., "Joint Video Multi-Frame Interpolation and Deblurring under Unknown Exposure Time," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Shang_Joint_Video_Multi-Frame_Interpolation_and_Deblurring_Under_Unknown_Exposure_Time_CVPR_2023_paper.html

[2] Z. Li *et al*., "Aggregating Nearest Sharp Features via Hybrid Transformers for Video Deblurring," *arXiv preprint arXiv:2309.07054*, 2024. [Online]. Available: https://arxiv.org/abs/2309.07054

[3] H. Zhao *et al*., "Swin-Diff: a Single Defocus Image Deblurring Network Based on Diffusion Model," *Complex & Intelligent Systems*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s40747-025-01789-w

[4] H. Chang *et al*., "Denoising Diffusion Models for Plug-and-Play Image Restoration," *arXiv preprint arXiv:2305.08995*, 2023. [Online]. Available: https://arxiv.org/abs/2305.08995

[5] X. Zhang *et al*., "Real-time Controllable Denoising for Image and Video," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_Real-Time_Controllable_Denoising_for_Image_and_Video_CVPR_2023_paper.pdf

[6] Y. Wang *et al*., "Towards High-Quality Real-Time Video Denoising with Pseudo Temporal Fusion Network," *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/372798480_Towards_High-Quality_Real-Time_Video_Denoising_with_Pseudo_Temporal_Fusion_Network

[7]Zhong, Z., Gao, Y., Zheng, Y., Zheng, B., & Sato, I. (2021). Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. International Journal of Computer Vision. https: //doi.org/10.1007/s11263-021-01510-0