Frames2Residual: Spatiotemporal Decoupling for Self-Supervised Video Denoising

Mingjie Ji[1], Zhan Shi[1], Kailai Zhou[2], Zixuan Fu[2], and Xun Cao[1,3,⋆]

[1] Nanjing University, Nanjing, China
[2] Nanyang Technological University, Singapore
[3] Key Laboratory of Optoelectronic Devices and Systems with Extreme Performances of MOE, Nanjing University, Nanjing, China
{jmj, shizhan}@smail.nju.edu.cn, {kailai.zhou, zixuan006}@ntu.edu.sg, caoxun@nju.edu.cn

Abstract. Self-supervised video denoising methods typically extend image-based frameworks into the temporal dimension, yet they often struggle to integrate inter-frame temporal consistency with intra-frame spatial specificity. Existing Video Blind-Spot Networks (BSNs) require noise independence by masking the center pixel, this constraint prevents the use of spatial evidence for texture recovery, thereby severing spatiotemporal correlations and causing texture loss. To address this, we propose Frames2Residual (F2R), a spatiotemporal decoupling framework that explicitly divides self-supervised training into two distinct stages: blind temporal consistency modeling and non-blind spatial texture recovery. In Stage 1, a blind temporal estimator learns inter-frame consistency using a frame-wise blind strategy, producing a temporally consistent anchor. In Stage 2, a non-blind spatial refiner leverages this anchor to safely reintroduce the center frame and recover intra-frame high-frequency spatial residuals while preserving temporal stability. Extensive experiments demonstrate that our decoupling strategy allows F2R to outperform existing self-supervised methods on both sRGB and raw video benchmarks. Code is available at https://github.com/m1NGGi/F2R.

Keywords: Video Denoising · Self-Supervised Learning

1 Introduction

Deep learning has achieved remarkable progress in image and video restoration. While image denoising leverages spatial self-similarity to suppress noise [11, 42], its efficacy is fundamentally constrained by the spatial entanglement of signal and noise [18]. To address this, video denoising introduces temporal dimension, utilizing correlations across frames to robustly identify coherent structures against random fluctuations [28, 31, 33]. However, the success of these methods predominantly hinges on supervised learning with paired datasets. Ground Truth (GT) is rarely available in practical applications like live-cell fluorescence

⋆ Corresponding author.

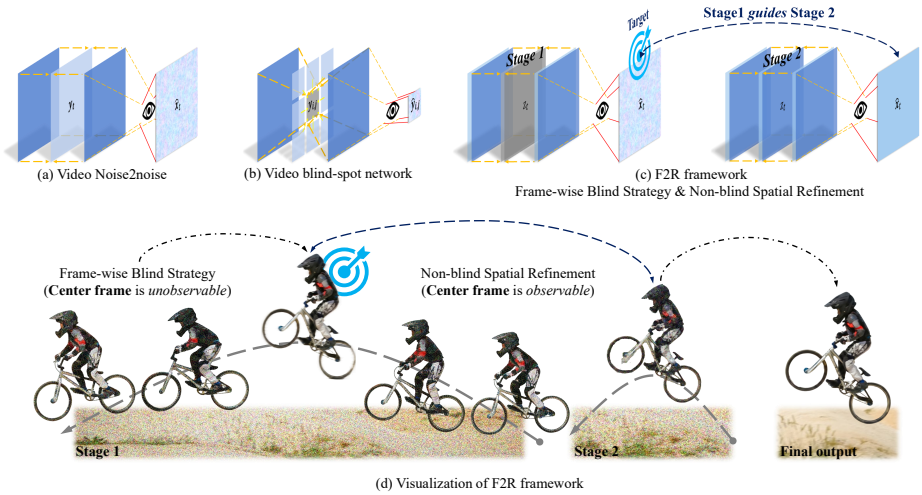

Fig. 1: Comparison of self-supervised video denoising paradigms during the inference phase. (a) Video Noise2Noise. Warping-based supervision violates noise independence and causes warping artifacts, leading to blurred details. (b) Video Blind-Spot Network. The inherent imposition of pixel discontinuities severs vital *spatiotemporal correlations*, resulting the loss in texture. (c) The proposed F2R framework. F2R employs joint spatial inputs $\{z_i\}$, where $z_i = \text{cat}(\hat{x}_i, r_i)$, $r_i = y_i - \hat{x}_i$, and $\hat{x}_i$ is derived from a pre-trained image denoiser $\mathcal{D}$. (d) Visualization of F2R framework. The gray dashed line illustrates the temporal consistency established by Stage 1. The final output acquires spatial specificity while maintaining temporal consistency, thereby effectively restoring *spatiotemporal correlations*.

microscopy or ultrafast transient imaging [26, 41]. This limitation necessitates self-supervised approaches that operate directly on raw noisy videos.

Most existing self-supervised video methods extend image-based frameworks, primarily Noise2Noise (N2N) [16] and Blind-Spot Networks (BSNs) [13,38] to the temporal dimension. Early Video N2N approaches [5,6,8] typically rely on motion compensation to align adjacent frames as supervision targets. Unfortunately, the interpolation inherent in warping inevitably disrupts pixel continuity and noise statistics [4], thereby fundamentally violating the independence assumption critical for self-supervision. Consequently, these methods often suffer from severe ghosting artifacts or over-smoothing in dynamic scenes (see Fig. 1(a)). To avoid this issue, Video BSNs frameworks [27] provide a statistically grounded alternative. By approximating the conditional expectation $\mathbb{E}[x_{t,i,j}|\Omega]$ solely from the spatiotemporal neighborhood Ω (see Fig. 1(b)), they circumvent the ill-posed motion alignment problem, ensuring that the training objective remains pristine and free from warping-induced corruptions. However, this strict exclusion strategy entails a specific trade-off. While this formulation strictly satisfies the noise independence assumption, the exclusion of direct spatial evidence leads to inherent pixel discontinuities, thereby disrupting continuous local variations [15, 35]. Consequently, this severs vital *spatiotemporal correlations*, defined as the struc-

tural continuity between a pixel and its local spatiotemporal neighborhood, ultimately leading to texture loss.

Simultaneously enforcing a blind-spot constraint for noise independence and utilizing direct spatial evidence for texture recovery are fundamentally conflicting objectives. To resolve this inherent conflict, we propose **Frames2Residual (F2R)**, a two-stage framework that fundamentally decouples temporal consistency modeling from spatial texture restoration. To enable this spatiotemporal decoupling and strictly focus the network on recovering high-frequency spatial textures lost by BSNs, we introduce a residual-domain learning formulation. Specifically, F2R leverages pre-computed image denoising outputs as a structural baseline (illustrated in Fig. 1(c)). By explicitly offloading static structural modeling to this prior, the model can concentrate exclusively on recovering spatial residuals. Similarly, to relieve the network from complex motion modeling during temporal aggregation, we introduce pre-computed optical flow. Directly applying pre-computed flow fields for explicit image warping inevitably corrupts vital details. Instead, we embed them into flow-guided implicit alignment modules, empowering the network to actively suppress resampling artifacts.

Under this residual-domain formulation, F2R utilizes the blindly estimated temporal consistency as a supervision target, creating a safe pathway to reintroduce direct spatial evidence for explicit texture recovery (illustrated in Fig. 1(d)). Specifically, Stage 1 trains a Blind Estimator (BE) via a frame-wise blind strategy. Because the target frame is unobservable, a Flow-Guided Attention Alignment Module (FAAM) replaces standard U-Net skip connections to stabilize cross-frame correlations and extract a temporally consistent anchor. Guided by this anchor, Stage 2 trains a non-blind Spatial Refiner (SR) via a recorruption strategy (substituting the original noisy center frame with a noise-injected anchor) to force active spatial texture recovery. With the target visible, a Flow-Guided Deformable Alignment Module (FDAM) leverages this spatial evidence to correct flow misalignments. By synthesizing complete spatiotemporal features, the SR precisely restores the missing spatial residuals absent in the structural baseline. Ultimately, this decoupled design unifies the inter-frame temporal consistency established in Stage 1 with the intra-frame spatial specificity recovered in Stage 2, successfully restoring the *spatiotemporal correlations*. Extensive evaluations on both sRGB and raw video benchmarks demonstrate that our proposed framework achieves state-of-the-art video denoising performance against current unsupervised methods.

In summary, our main contributions are:

- We identify an inherent conflict in self-supervised video denoising: enforcing noise independence through blind-spot prediction prevents the model from exploiting direct spatial evidence for texture recovery. To resolve this issue, we propose F2R, a two-stage framework that decouples temporal consistency modeling from spatial texture restoration.
- To enable this decoupling, we introduce a residual-domain learning formulation that leverages image denoising priors as a structural baseline, allowing the model to focus on recovering high-frequency spatial residuals while effectively aggregating temporal information.

– Our approach surpasses current unsupervised methods on both raw and sRGB video denoising benchmarks, demonstrating its superior effectiveness.

2 Related Work

2.1 Supervised Video Denoising

Supervised video denoising typically relies on GT sequences to learn spatiotemporal features [1,2,17,19,20,23,31,33,34,39,41]. TOFlow [39] and DVDNet [30] explicitly computed optical flow to align neighboring frames. To avoid artifacts from flow estimation errors, FastDVDNet [31] proposed a cascade of U-Nets without explicit motion compensation. EDVR [34] and RViDeNet [41] utilize Deformable Convolution Networks (DCN) [34] for feature-level alignment, while Transformer-based models like VRT [19], RVRT [20], and Tempformer [28] leverage attention mechanisms for long-range temporal modeling. Despite their high performance, the heavy reliance on paired clean-noisy data severely restricts their applicability in real-world scenarios.

2.2 Self-supervised Video Denoising

In the absence of GT, self-supervised video denoising has emerged as a vital paradigm to circumvent the dependence on paired training data. Early self-supervised approaches [5,6,8,40] extend the N2N assumption to video by warping adjacent frames to construct clean targets. However, the interpolation and resampling inherent in warping inevitably alter noise statistics and introduce artifacts, potentially violating the independence assumption required for N2N training. Alternatively, some recent works [9, 14, 43] adopt recorruption or pseudo-impairment schemes to construct self-supervised targets without explicit warping. To better preserve statistical validity, Video BSN approaches [36,43] employ blind-spot constraints, predicting the center pixel solely from its spatiotemporal context. Based on this, implicit alignment methods like UDVD [27] avoid explicit flow artifacts but rely on frame stacking, leading to limited temporal receptive fields. To improve temporal efficiency, recurrent variants [4, 36] leverage temporal recurrence to reuse hidden states. However, their sequential processing risks overfitting to training sequences. Crucially, these methods often sacrifice local detail and introduce pixel discontinuities due to strict noise independence assumptions, leading to an inevitable loss of *spatiotemporal correlations* that fundamentally restricts the restoration of fine-grained spatial textures.

3 Methodology

To resolve the inherent conflict between enforcing noise independence and utilizing spatial evidence for texture recovery, we propose **F2R**, a framework centered on a *spatiotemporal decoupling* strategy. Specifically, F2R formulates the training process as a two-stage decomposition problem to effectively address the

conflicting objectives inherent in video BSNs. By leveraging pre-computed image denoising outputs to offload static structural modeling, F2R focuses exclusively on recovering the remaining residuals. In Stage 1, a Blind Estimator extracts temporal consistency without using spatial evidence. Subsequently in Stage 2, a Spatial Refiner safely reintegrates spatial evidence, synthesizing the complete spatiotemporal information to restore the high-frequency residuals that the image denoiser missed. Both models in this strategy share an identical standard 4-level U-Net backbone, differing exclusively in their skip connections. Specifically, the Blind Estimator employs FAAM, whereas the Spatial Refiner utilizes FDAM. Notably, this dual-model decoupling is exclusive to training; during inference, only the trained Spatial Refiner is deployed.

3.1 Stage 1: Blind Temporal Estimating via Frame-wise Blind Strategy

Given a noisy video sequence $\mathcal{Y} = \{y_1, \ldots, y_T\}$ where $y_i = x_i + n_i$, the objective of Stage 1 is to isolate the temporal component by constructing a BE tasked with estimating the clean center frame x_t exclusively from inter-frame temporal consistency. To achieve this, we introduce a **Frame-wise Blind Strategy**, as shown in Fig. 2(b). Specifically, we exclude the entire center frame y_t from the input, defining the temporal context as $\Omega_{\text{temporal}} = \{y_i\}_{i \neq t}$. This explicit exclusion naturally satisfies the noise independence assumption, forcing the network to approximate the conditional expectation $\mathbb{E}[x_t|\Omega_{\text{temporal}}]$. By blocking the spatial pathway, we ensure the learned features represent pure temporal consistency, decoupled from intra-frame noise and texture.

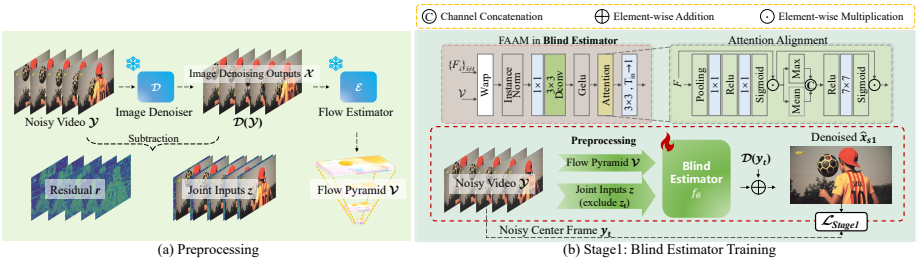

Fig. 2: Frame-wise Blind Strategy. (a) Preprocessing for input construction. (b) Blind estimator training. The flow pyramid $\mathcal{V}$ is constructed by downsampling the base flow $\{\mathcal{V}_{i\to t}\}_{i\neq t}$ and halving its magnitude at each level.

Residual-Domain Optimization with Image Denoiser. Directly regressing the clean frame x_t from temporal neighbors is ill-posed due to complex motions and occlusions. To resolve this, we shift the learning target to the residual domain by leveraging pre-computed image denoising outputs $\mathcal{X}$. A pre-trained image denoiser $\mathcal{D}$ is applied frame-wise to produce the image denoising output $\hat{x}_i = \mathcal{D}(y_i)$.

This process yields a deterministic structural baseline $\hat{x}_i$ and a corresponding high-frequency residual $r_i = y_i - \hat{x}_i$. By offloading the modeling of static spatial structures to the image denoiser, this explicit decomposition forces the network to focus exclusively on isolating and recovering the high-frequency texture residual. For each neighboring frame, we construct a joint input $z_i = \mathrm{cat}(\hat{x}_i, r_i)$. The inclusion of $\{\hat{x}_i\}_{i\neq t}$ offers a consistent structural reference to reduce noise ambiguity, while $\{r_i\}_{i\neq t}$ supplies essential high-frequency cues. The process is illustrated in Fig. 2(a).

Successfully extracting accurate temporal residuals from these joint inputs requires the network to first overcome the complexities of temporal feature alignment. To prevent the network from expending its capacity on complex motion modeling, we introduce an architectural decoupling to separate spatial extraction from temporal fusion. Specifically, a standard U-Net backbone processes each frame strictly independently to extract intra-frame features, intentionally deferring all inter-frame temporal fusion to the skip connections. Within these connections, we employ pre-computed optical flow to guide implicit feature alignment. This design effectively offloads the alignment burden while empowering the network to actively suppress resampling artifacts inherent to explicit image warping. However, in this blind stage, the absence of the center frame y_t makes aggressive geometric warping ill-posed, as predicting precise offsets without a valid reference often leads to structural distortions.

Therefore, we replace standard U-Net skip connections with a FAAM to implement a conservative filtration strategy. As shown in Fig. 2(b), FAAM initially utilizes explicit optical flow $\{\mathcal{V}_{i\to t}\}_{i\neq t}$ (derived from $\mathcal{X}$ using a pre-trained flow estimator $\mathcal{E}$ to mitigate noise interference) to warp neighboring features $\{F_i\}_{i\neq t}$ to the center. To mitigate resulting resampling artifacts, the warped features are stacked and processed through a sequential dual-attention mechanism [10, 37]. A channel attention module acts as a temporal selector to dynamically reweight frames and suppress unreliable motion, followed by a spatial attention module that highlights valid local regions to filter out warping distortions. Finally, a 1×1 convolution aggregates these refined features. This distills reliable inter-frame consensus, focusing the BE f_θ on robust temporal residual extraction. The optimization objective is:

$$\mathcal{L}_{\text{Stage1}} = \mathbb{E}\left\| f_\theta(\{z_i\}_{i\neq t}, \mathcal{V}) - r_t \right\|. \tag{1}$$

Here, the noisy residual $r_t = y_t - \mathcal{D}(y_t)$ serves as the supervision target, while the network strictly observes only the temporal neighbors $\{z_i\}_{i\neq t}$. The resulting estimate $\hat{x}_{s1} = \hat{x}_t + f_\theta(\{z_i\}_{i\neq t}, \mathcal{V})$ represents a temporally consistent estimate. However, since this residual is derived exclusively from the decoupled temporal consistency, it inherently lacks frame-specific, non-redundant texture residuals. This intrinsic limitation necessitates a subsequent stage for spatial refinement.

3.2 Stage 2: Non-blind Spatial Refinement via Recorruption Strategy

Stage 1 successfully isolates temporal consistency, yielding a temporally consistent estimate $\hat{x}_{s1}$. To complete our spatiotemporal decoupling framework, Stage 2 must reintegrate the specific high-frequency spatial textures that the image denoiser $\mathcal{D}$ originally smoothed away, without compromising the temporal stability achieved in Stage 1. To this end, we introduce a **Non-blind Spatial Refinement** phase driven by a **Recorruption** strategy (Fig. 3(a)). This strategy creates a safe pathway to reintroduce direct spatial evidence.

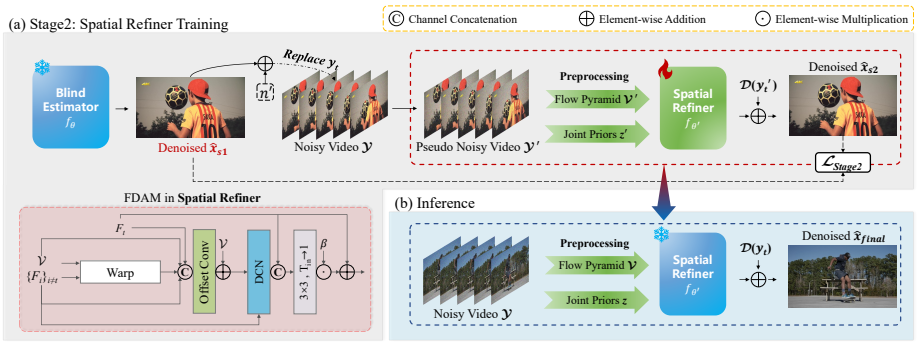

Fig. 3: (a) Training and (b) Inference phases of the Spatial Refiner. Notably, the preprocessing operations are strictly identical across both phases. The recorruption noise n' is sampled from the known noise model.

We leverage the fixed mapping of the pre-trained image denoiser $\mathcal{D}$. Since $\mathcal{D}$ systematically suppresses specific high-frequency textures alongside noise, we can train a non-blind SR, $f_{\theta'}$, to predict this deterministic spatial gap. We designate $\hat{x}_{s1}$ as our "temporal anchor". To construct the training target, we directly synthesize a pseudo noisy sequence $\mathcal{Y}'$ by replacing the original noisy center frame y_t in $\mathcal{Y}$ with a recorrupted temporal anchor $y'_t = \hat{x}_{s1} + n'$, where n' is from a known noise model. The neighboring noisy frames remain strictly unchanged. Injecting this noise n' exclusively into the temporal anchor effectively masks the temporal residual embedded in $\hat{x}_{s1}$, compelling the SR to perform active denoising rather than learning a trivial identity mapping.

We reapply the frozen denoiser $\mathcal{D}$ to the recorrupted center frame y'_t. The missing residual is defined as the difference between our temporal anchor and the new image denoising output:

$$r'_t = \hat{x}_{s1} - \mathcal{D}(y'_t), \tag{2}$$

This precisely quantifies the deterministic texture loss incurred by $\mathcal{D}$.

In stark contrast to Stage 1, the SR $f_{\theta'}$ is non-blind and has full access to the unmasked center frame y'_t. The updated joint inputs for the center frame

become $z'_t = \text{cat}(\mathcal{D}(y'_t), y'_t - \mathcal{D}(y'_t))$, while neighboring inputs $\{z_i\}_{i\neq t}$ remain identical to Stage 1. This visibility provides the direct spatial evidence necessary for precise calibration of geometric misalignments, crucial for recovering high-frequency spatial textures. Consequently, we employ a FDAM to achieve aggressive, sub-pixel alignment.

FDAM uses DCN [9, 34] to handle complex, non-rigid motions missed by pre-computed optical flow. To stabilize DCN optimization, we leverage the explicit optical flow $\{\mathcal{V}'_{i\to t}\}_{i\neq t}$ (derived from the updated image denoising sequence $\mathcal{D}(\mathcal{Y}')$) as a base offset initialization [2]. The module concatenates the reference feature F_t, neighboring features $\{F_i\}_{i\neq t}$, warped neighboring features, and the flow field to regress residual offsets. This decouples global motion from local deformations, focusing DCN on fine-grained sub-pixel adjustments. The aligned features are fused and added back to the reference via a learnable scale parameter β, ensuring the module can utilize temporal information to actively compensate for the spatial residual extraction process. The non-blind SR is trained to minimize:

$$\mathcal{L}_{\text{Stage2}} = \mathbb{E}\left\| f_{\theta'}(\{z'\}, \mathcal{V}') - r'_t \right\|, \tag{3}$$

By minimizing this loss, the SR $f_{\theta'}$ learns to fuse the visible intra-frame spatial cues with the broader temporal context, reclaiming the spatial specificity absent from the Stage 1 output. Because $\mathcal{D}$ acts as a consistent spatial filter, the degradation patterns it introduces are identical across both synthetic recorrupted training and actual inference, making the learned residual mapping directly transferable.

As illustrated in Fig. 3(b), during inference, the BE f_θ is discarded. The converged non-blind SR $f_{\theta'}$ is applied directly to the original noisy joint inputs $\{z\}$ and optical flow $\mathcal{V}$ to predict the final residual:

$$\hat{x}_{\text{final}} = f_{\theta'}(\{z\}, \mathcal{V}) + \mathcal{D}(y_t). \tag{4}$$

This final output $\hat{x}_{\text{final}}$ seamlessly unifies the inter-frame temporal consistency established in Stage 1 with the intra-frame spatial specificity recovered in Stage 2, successfully reconstructing the vital *spatiotemporal correlations*.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the proposed F2R on two distinct video denoising tasks: synthetic Gaussian denoising and real-world raw video denoising. For synthetic Gaussian denoising, we utilize the DAVIS 2017 [24] (480P) training set for model training. To generate noisy video sequences, Additive White Gaussian Noise (AWGN) with a standard deviation σ varying from 10 to 50 uniformly is added to the clean videos. For evaluation, we employ the DAVIS validation set and the Set8 [31] dataset. For real raw video denoising, we employ the CRVD [41] dataset. Due to the limited size of the dataset, we follow the standard unsupervised

protocol where all self-supervised models are trained directly on the noisy videos of the CRVD indoor set without accessing ground truth labels.

Implementation Details. Our framework leverages robust pre-trained priors to ensure both efficiency and performance. For the image denoiser, we employ NAFNet [3], following the training strategy in TAP [9]. For the optical flow estimator, we adopt the pre-trained PWC-Net [29], following the standard practices in [7, 25, 32]. During training, we randomly crop patches of size 256×256 with a batch size of 4. For synthetic Gaussian denoising, the input temporal window size is set to $T = 9$, while for real raw video denoising, we set $T = 7$. Raw videos are packed into RGBG sequences during training. The feature channel number of the network is fixed at $C = 72$, and the network is optimized using the Adam optimizer [12] with L_1 loss for synthetic Gaussian denoising and L_2 loss for real noise. We train for 200K iterations, with the learning rate initialized at 3×10^{-4} and gradually decayed using a cosine annealing strategy [21]. To ensure statistical consistency in Stage 2, for synthetic experiments, we sample n' from $\mathcal{N}(0, \sigma^2)$ matching the specific noise level σ of the current model. In contrast, for real-world tasks where a unified model is trained across all ISO levels, we randomly sample signal-dependent noise based on the calibrated noise profile corresponding to the ISO of the input video. All experiments are implemented in PyTorch and conducted on an NVIDIA L20 GPU. We employ Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as evaluation metrics.

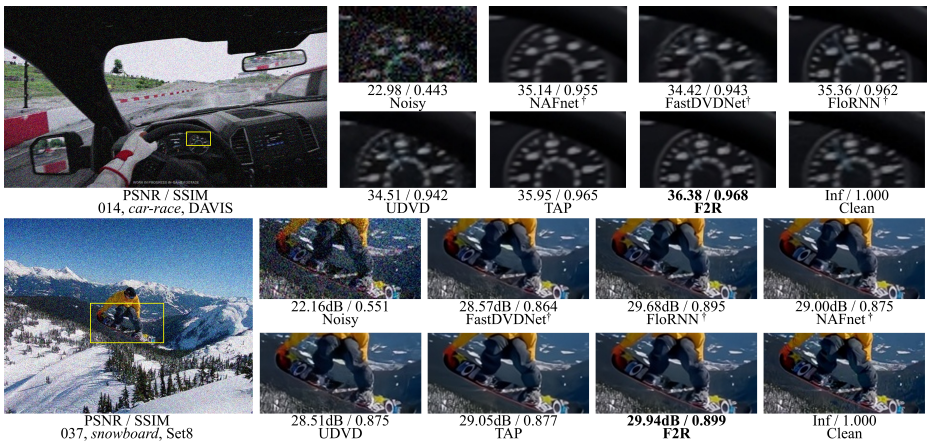

Fig. 4: Visual comparison on DAVIS (top) and Set8 (bottom) datasets under noise level $\sigma = 30$. We compare our F2R with supervised (FastDVDNet, FloRNN, NAFNet) and unsupervised (UDVD, TAP) methods, with PSNR/SSIM metrics shown below each patch. † indicates the supervised method.

4.2 Comparison with State-of-the-Art

The experimental results for competing methods are either directly cited from their original papers or reproduced by running the official code. In cases where the code is unavailable, the corresponding results are omitted. We compare our

proposed F2R with state-of-the-art self-supervised video denoising methods, i.e., MF2F [6], RFR [14], UDVD [43], RDRF [36], ER2R [43], and TAP [9]. For a comprehensive comparison, we also include traditional and supervised video denoising methods such as VBM4D [22], FastDVDNet [31], PaCNet [33], and FloRNN [17]. Note that methods utilizing the same video sequence for both training and testing are tagged with 'T'.

Table 1: Quantitative comparison (PSNR/SSIM) on DAVIS and Set8 datasets under different noise levels. The best results for unsupervised methods are **bolded**, and the second best are underlined.

Dataset		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 [22]	37.58	-	33.88	-	31.65	-	30.05	-	28.80	-	32.39	-
	Supervised	NAFNet [3]	38.79	0.965	35.37	0.933	33.47	0.904	32.17	0.879	31.18	0.858	34.20	0.908
		FastDVD [31]	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 [33]	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 [17]	40.16	0.976	37.52	0.956	35.89	0.944	34.66	0.929	33.67	0.913	36.38	0.944
	Self-supervised	MF2F-T [6]	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-T [14]	39.31	-	36.15	-	34.28	-	32.92	-	31.86	-	34.90	-
		UDVD [43]	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 [36]	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-T [43]	39.52	-	36.49	-	34.60	-	33.29	-	32.25	-	35.23	-
		TAP-T [9]	39.80	0.973	36.74	0.950	34.84	0.926	33.49	0.905	32.49	0.886	35.48	0.928
		F2R (Ours)	**40.25**	**0.976**	**37.33**	**0.958**	**35.56**	**0.940**	**34.30**	**0.923**	**33.28**	**0.906**	**36.14**	**0.941**
Set8	Traditional	VBM4D [22]	36.05	-	32.19	-	30.00	-	28.48	-	27.33	-	30.81	-
	Supervised	NAFNet [3]	36.52	0.943	33.55	0.802	31.81	0.869	30.59	0.842	29.65	0.818	32.43	0.875
		FastDVD [31]	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 [33]	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 [17]	37.57	0.964	34.67	0.938	32.97	0.914	31.75	0.891	30.80	0.870	33.55	0.915
	Self-supervised	MF2F-T [6]	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-T [14]	36.77	-	33.64	-	31.82	-	30.52	-	29.50	-	32.45	-
		UDVD [43]	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 [36]	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-T [43]	37.55	-	34.34	-	32.45	-	31.09	-	30.05	-	33.10	-
		TAP-T [9]	**38.02**	**0.958**	35.07	0.927	33.42	0.900	32.10	0.875	31.16	0.852	33.95	0.902
		F2R (Ours)	37.85	**0.958**	**35.38**	**0.935**	**33.82**	**0.913**	**32.68**	**0.894**	**31.75**	**0.875**	**34.30**	**0.915**

Experiments on synthetic Gaussian denoising. Following standard protocols [31], we evaluate AWGN removal on the DAVIS and Set8 benchmarks. As detailed in Tab. 1, our F2R consistently outperforms unsupervised methods by large margins. On DAVIS, F2R achieves the highest average PSNR of 36.14 dB, surpassing the recent competitor TAP-T [9] by 0.66 dB and RDRF [36] by 0.96 dB. Notably, on the Set8 dataset, F2R (34.30 dB) not only outperforms unsupervised baselines but also surpasses the supervised FloRNN [17] by 0.75 dB, narrowing the gap between supervised and unsupervised paradigms. Fig. 4 illustrates video denoising examples on DAVIS and Set8 datasets. In the challenging *car-race* scene where even the clean reference is motion-blurred, F2R leverages complementary temporal cues to recover sharper dashboard scale markings and clearer structures than competing methods. More critically, in the dynamic texture region on the pants of the snowboarder, TAP fails to preserve the fabric structure despite incorporating image denoising outputs, indicating suboptimal utilization. Conversely, F2R effectively recovers these high-frequency spatial textures, validating the effectiveness of our strategy.

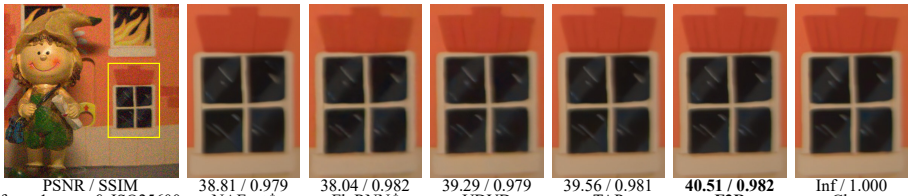

Fig. 5: Visual comparison on CRVD indoor dataset. The results have been converted to the sRGB domain with the pretrained ISP provided in [41] for visualization. † indicates the supervised method.

Table 2: Quantitative comparison on CRVD indoor dataset under different ISO levels. The best results among unsupervised methods are **bolded**.

	Method \ ISO	1600	3200	6400	12800	25600	Average
Supervised	NAFNet [3]	48.02	46.05	44.04	41.44	41.49	44.21
	RViDeNet [41]	47.74	45.91	43.85	41.20	41.17	43.97
	FloRNN [17]	48.81	47.05	45.09	42.63	42.19	45.15
Unsupervised	UDVD [43]	48.02	46.44	44.74	42.21	42.13	44.71
	RDRF [36]	48.38	46.86	45.24	42.72	42.25	45.09
	TAP [9]	48.85	47.03	45.11	42.44	42.33	45.15
	F2R (Ours)	**49.41**	**47.52**	**45.66**	**43.03**	**42.82**	**45.69**

Experiments on real raw video denoising. Tab. 2 reports the evaluation results on the CRVD indoor dataset. F2R still outperforms the leading unsupervised scheme TAP [9] by 0.54 dB and surpasses the supervised competitor FloRNN [17], demonstrating superior generalization in complex real-world noise distributions. To validate the perceptual quality, we provide a visual comparison in Fig. 5. In this challenging low-light indoor scene, preserving the structural integrity of the window panes is critical. Existing methods exhibit characteristic failure modes. The existing Video BSN method, UDVD [43], fails to retain the sharp edges of the window frame, causing severe blurring. Similarly, the spatial-only baseline NAFNet [3] tends to aggressively suppress high-frequency content, leading to over-smoothed textures. Although TAP [9] leverages temporal information to compensate for spatial deficits, it still fails to fully reconstruct texture details in high-frequency regions. In contrast, F2R effectively reconstructs clear structural lines and authentic wall textures, yielding a result that is visually closest to the clean reference.

Fig. 6: Visual ablation of intermediate stages in F2R. Evaluated on the fast-motion *motorbike* sequence from Set8 with noise level $\sigma = 30$. **Stage 2** represents the result of cascaded training of Stage 1 and Stage 2, as denoised by the $f_{\theta'}$.

4.3 Ablation Study

Impact of Individual Stages in F2R. We validate the necessity of our dual-stage design in Tab. 3. Compared to the baseline, Stage 1 alone yields significant improvements, confirming its ability to estimate stable temporal residual. Conversely, applying Stage 2 independently fails. Without the temporal anchor from Stage 1, the network simply learns a trivial identity mapping of the image denoising output. This proves that Stage 2 is strictly contingent upon the temporal consistency established by Stage 1. Ultimately, the synergy of both stages achieves peak performance, providing further significant gains beyond Stage 1 alone on the CRVD dataset. As visualized in Fig. 6, the image denoising output (NAFnet) suffers from severe blur. While Stage 1 stabilizes the background, it omits unique spatial texture, leaving the "MOTOREX" and "RK" words indistinct due to the intrinsic blindness of the inference process. Stage 2 remedies this via spatial refinement, utilizing the visible center frame to recover sharp, legible text. This validates the effective decomposition of F2R, where Stage 1 secures temporal consistency while Stage 2 refine spatial specificity.

Table 3: Impact of individual stages in F2R. We investigate the contribution of Stage 1 and Stage 2. For the Stage 2-only setting, we directly utilize the initial joint inputs to perform re-corruption. Note that the identical scores in the first and third columns are due to rounding.

Component	Methods			
Per-denoised	✓	✓	✓	✓
Stage 1		✓		✓
Stage 2			✓	✓
	DAVIS in $\sigma = 30$			
PSNR (dB)	33.47	34.36	33.47	**35.56**
SSIM	0.904	0.930	0.904	**0.940**
	Set8 in $\sigma = 30$			
PSNR (dB)	31.81	32.70	31.81	**33.82**
SSIM	0.869	0.903	0.869	**0.913**
	CRVD in ISO = 25600			
PSNR (dB)	41.49	41.88	41.49	**42.82**

Effectiveness of key components in F2R. We validate our design in Tab. 4. The baseline, a standard 4-level U-Net that adopts channel-wise sequence stacking and direct skip connections, suffers from inefficient optimization and feature misalignment arising from implicit motion modeling. Replacing direct skips with FAAMs decouples spatiotemporal features, mitigating alignment issues and obtaining 3.16 dB gain. Most significantly, adopting the joint inputs yields a massive 8.53 dB gain. This confirms that shifting optimization to the residual domain drastically simplifies learning by focusing exclusively on residuals. Ultimately, the synergy of explicit guidance and residual optimization achieves peak performance, validating their strong complementarity.

Table 4: Effectiveness of key components. The baseline is a standard 4-level U-Net with channel-wise frame concatenation and direct skip connections. We evaluate replacing direct skips with FAAM ($\mathcal{V}$) and shifting raw inputs to joint inputs (z) for residual prediction. Evaluated in Stage 1 on DAVIS ($\sigma = 30$).

Component	Methods			
Flow-guided ($\mathcal{V}$)		✓		✓
Joint inputs (z)			✓	✓
PSNR (dB)	25.06	28.22	33.59	**34.36**
SSIM	0.731	0.870	0.907	**0.930**

Necessity of Decoupled Alignment Modules. Tab. 5 verifies that alignment modules adapt to specific objective of each stage. An uniform FAAM (35.37 dB) ensures stability but lacks the sub-pixel precision required for Stage 2 spatial recovery. Conversely, an uniform FDAM (35.53 dB) improves overall alignment but remains suboptimal in Stage 1, as blind estimation lacks the center reference needed to reliably guide complex deformable offsets. By decoupling these modules and utilizing the robust FAAM for Stage 1 blind extraction alongside the precise FDAM for Stage 2 non-blind refinement, F2R achieves peak performance.

Table 5: Validation of strategy-specific flow-guided implicit alignment modules on the DAVIS dataset ($\sigma = 30$).

Stage 1	Stage 2	PSNR (dB)	SSIM
FAAM	FAAM	35.37	0.939
FDAM	FDAM	35.53	0.940
FAAM	**FDAM**	**35.56**	**0.940**

Influence of Temporal Window Size. To strictly isolate temporal estimating from spatial refinement, we ablate the window size $T \in \{3, 5, 7, 9, 11\}$ exclusively on the Stage 1 intermediate output ($\hat{x}_{s1}$) using synthetic datasets ($\sigma = 30$). As Tab. 6 shows, while increasing T initially boosts performance, it quickly exhibits diminishing returns. Expanding from $T = 9$ to $T = 11$ yields negligible gains (e.g., +0.03 dB on DAVIS) while significantly exacerbating computational overhead and the risk of long-range flow alignment artifacts. Thus, $T = 9$ serves as the optimal trade-off for synthetic datasets. For the real-world CRVD dataset, we utilize $T = 7$, naturally constrained by its maximum captured sequence length.

Fair Comparison under Identical Temporal Window. To ensure a fair comparison, we evaluate F2R under the strict 5-frame setting ($T = 5$) adopted by baselines such as FastDVDNet [31], UDVD [43], and TAP-T [9]. As reported in Tab. 7, F2R achieves 35.19 dB on the DAVIS dataset ($\sigma = 30$) in this restricted context. This outperforms the leading method, TAP-T, by 0.35 dB, confirming the superiority of F2R.

Table 6: Ablation on the temporal window size T. Results report PSNR (dB) on the Stage 1 intermediate output using synthetic datasets ($\sigma = 30$). Values in parentheses indicate the marginal gain over the previous T.

Dataset	Temporal Window Size T				
	$T = 3$	$T = 5$	$T = 7$	$T = 9$	$T = 11$
DAVIS	33.96	34.23 (+0.27)	34.31 (+0.08)	**34.36 (+0.05)**	34.39 (+0.03)
Set8	32.30	32.56 (+0.26)	32.65 (+0.09)	**32.70 (+0.05)**	32.74 (+0.04)

Table 7: Fair comparison under an identical temporal window ($T = 5$). Evaluated on the DAVIS dataset under the noise level $\sigma = 30$. All methods strictly utilize 5 input frames. Best results are highlighted in **bolded**.

Method	Type	Frames (T)	PSNR (dB)	SSIM
FastDVDNet [31]	Supervised	5	34.04	0.917
UDVD [43]	Self-supervised	5	34.09	0.918
TAP-T [9]	Self-supervised	5	34.84	0.926
F2R (Ours)	Self-supervised	5	**35.19**	**0.934**

5 Conclusion

In this paper, we propose F2R, a framework empowered by a *spatiotemporal decoupling* strategy, designed to address the critical bottleneck of severed *spatiotemporal correlations* in existing Video BSNs. Fundamentally, F2R operates through two tailored self-supervised strategies. It first employs a frame-wise blind strategy to robustly estimate a temporally consistent anchor. Subsequently, a recorruption strategy leverages this anchor and reintroduces the excluded frame to safely recover the high-frequency spatial residuals discarded by the image denoiser. This decoupled design ensures that the recovered spatial specificity strictly maintains the established temporal stability, effectively reconstructing the *spatiotemporal correlations*. Comprehensive experiments on sRGB and raw video datasets verify that F2R achieves state-of-the-art performance.

Acknowledgements This work was supported by the National Key Research and Development Program of China (2023YFF0713300).

References

1. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: 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 (CVPR). pp. 4947–4956 (2021)
2. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5972–5981 (2022)

3. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 17–33. Springer (2022)
4. Chen, Z., Jiang, T., Hu, X., Zhang, W., Li, H., Wang, H.: Spatiotemporal blind-spot network with calibrated flow alignment for self-supervised video denoising. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 2411–2419 (2025)
5. Davy, A., Ehret, T., Morel, J.M., Arias, P., Facciolo, G.: A non-local cnn for video denoising. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 2409–2413. IEEE (2019)
6. Dewil, V., Anger, J., Davy, A., Ehret, T., Facciolo, G., Arias, P.: Self-supervised training for blind multi-frame video denoising. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2724–2734 (2021)
7. Dong, Q., Fu, Y.: Memflow: Optical flow estimation and prediction with memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19068–19078 (2024)
8. Ehret, T., Davy, A., Morel, J.M., Facciolo, G., Arias, P.: Model-blind video denoising via frame-to-frame training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11369–11378 (2019)
9. Fu, Z., Guo, L., Wang, C., Wang, Y., Li, Z., Wen, B.: Temporal as a plugin: Unsupervised video denoising with pre-trained image denoisers. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–367. Springer (2024)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018)
11. Huang, T., Li, S., Jia, X., Lu, H., Liu, J.: Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14776–14785. IEEE, Nashville, TN, USA (2021)
12. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2129–2137 (2019)
14. Lee, S., Cho, D., Kim, J., Kim, T.H.: Restore from restored: Video restoration with pseudo clean video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3537–3546 (2021)
15. Lee, W., Son, S., Lee, K.M.: Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17725–17734 (2022)
16. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv:1803.04189 (2018)
17. Li, J., Wu, X., Niu, Z., Zuo, W.: Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 592–609. Springer (2022)
18. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)

19. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. IEEE Transactions on Image Processing **33**, 2171–2182 (2024)
20. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems (NeurIPS) **35**, 378–393 (2022)
21. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
22. Maggioni, M., Boracchi, G., Foi, A., Egiazarian, K.: Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. IEEE Transactions on Image Processing **21**(9), 3952–3966 (2012)
23. Maggioni, M., Huang, Y., Li, C., Xiao, S., Fu, Z., Song, F.: Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3466–3475 (2021)
24. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
25. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4161–4170 (2017)
26. Sadri, A., Petersen, T.C., Terzoudis-Lumsden, E.W., Esser, B.D., Etheridge, J., Findlay, S.D.: Unsupervised deep denoising for four-dimensional scanning transmission electron microscopy. NPJ Computational Materials **10**(1), 243 (2024)
27. Sheth, D.Y., Mohan, S., Vincent, J.L., Manzorro, R., Crozier, P.A., Khapra, M.M., Simoncelli, E.P., Fernandez-Granda, C.: Unsupervised deep video denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1759–1768 (2021)
28. Song, M., Zhang, Y., Aydın, T.O.: Tempformer: Temporally consistent transformer for video denoising. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 481–496. Springer (2022)
29. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8934–8943 (2018)
30. Tassano, M., Delon, J., Veit, T.: Dvdnet: A fast network for deep video denoising. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 1805–1809. IEEE (2019)
31. Tassano, M., Delon, J., Veit, T.: Fastdvdnet: Towards real-time deep video denoising without flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1354–1363 (2020)
32. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 402–419. Springer (2020)
33. Vaksman, G., Elad, M., Milanfar, P.: Patch craft: Video denoising by deep modeling and patch matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2157–2166 (2021)
34. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 0–0 (2019)

35. Wang, Z., Liu, J., Li, G., Han, H.: Blind2unblind: Self-supervised image denoising with visible blind spots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2027–2036 (2022)
36. Wang, Z., Zhang, Y., Zhang, D., Fu, Y.: Recurrent self-supervised video denoising with denser receptive field. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 7363–7372 (2023)
37. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
38. Wu, X., Liu, M., Cao, Y., Ren, D., Zuo, W.: Unpaired learning of deep image denoising. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 352–368. Springer (2020)
39. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
40. Yu, S., Park, B., Park, J., Jeong, J.: Joint learning of blind video denoising and optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 500–501 (2020)
41. Yue, H., Cao, C., Liao, L., Chu, R., Yang, J.: Supervised raw video denoising with a benchmark dataset on dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2301–2310 (2020)
42. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient Transformer for High-Resolution Image Restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5718–5729. IEEE, New Orleans, LA, USA (2022)
43. Zheng, H., Pang, T., Ji, H.: Unsupervised deep video denoising with untrained network. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 3651–3659 (2023)