



Unsupervised Deep Video Denoising

Dev Yashpal Sheth, ICCV, 2021

Saebom Lee



Abstract

In many applications, such as microscopy, noiseless videos are not available.



To address this, we propose an Unsupervised Deep Video Denoiser, a CNN architecture designed to be trained exclusively with noisy data.



In contrast to many current approaches to video denoising, UDVD does not require explicit motion compensation.



Thus, the network learns to perform implicit motion compensation, even though it is only trained for denoising.



Contents

Sec1. Introduction

Sec2. Background and Related Work

Sec3. Unsupervised Deep Video Denoising

Sec4. Datasets

Sec5. Experiments and Results

Sec6. Automatic Motion Compensation

Sec7. Conclusion



Sec1. Introduction



Introduction

Paragraph 01

Convolutional neural networks are typically trained using a database of clean videos, which are corrupted with simulated noise. However, in applications such as microscopy, noiseless ground truth videos are often not available. To address this issue, we propose a method to train a video denoising CNN without access to supervised data, which we call Unsupervised Deep Video Denoising (UDVD). Here, we propose a blind-spot architecture that processes the surrounding spatio-temporal neighborhood to denoise videos.

Paragraph 02

When combined with aggressive data augmentation and early stopping, it can produce high quality denoising even when trained exclusively on a single brief noisy video sequence.

In contrast, we demonstrate that UDVD can effectively denoise three different real-world datasets: raw videos from surveillance cameras, fluorescence-microscopy videos of cells, and electron-microscopy videos of catalytic nanoparticles.



Introduction

Paragraph 03

Nearly all existing approaches to video denoising use estimates of optical flow to adaptively compensate for the motion of objects in the video. UDVD yield excellent empirical performance without explicit estimation of optical flow. Instead, we use a gradient-based analysis to show that both UDVD and supervised CNNs perform spatio-temporal adaptive filtering, which is aligned with underlying motion. Thus, these CNNs are automatically performing implicit motion compensation.



Sec2. Background and Related Work



Background and Related Work

- Traditional and CNN-based video denoising
 - Traditional techniques for single image denoising include nonlinear filtering, sparse prior methods, and nonlocal means. In order to exploit the spatio temporal structure of the video, these methods typically employ motion compensation based on estimates of optical flow.
 - The CNNs are trained to minimize the mean squared error between the network output and ground truth using large databases of natural images/videos.



Background and Related Work

- Video denoising without motion compensation
 - Three recent methods perform video denoising without explicit motion estimation .
 - VNLnet
 - ViDeNN
 - FastDVDnet
 - In this work we show that such CNNs actually performs implicit motion estimation, which can be revealed through a gradient-based analysis.



Background and Related Work

- unsupervised denoising
 - Using the N2N framework to perform unsupervised video denoising requires warping adjoining frames, which in turn requires explicit motion compensation, and accurate occlusion estimation.
 - In order to bypass these issues, we develop a blind-spot network that trains denoising CNNs by fitting the noisy data directly.



Sec3. Unsupervised Deep Video Denoising

Unsupervised Deep Video Denoising

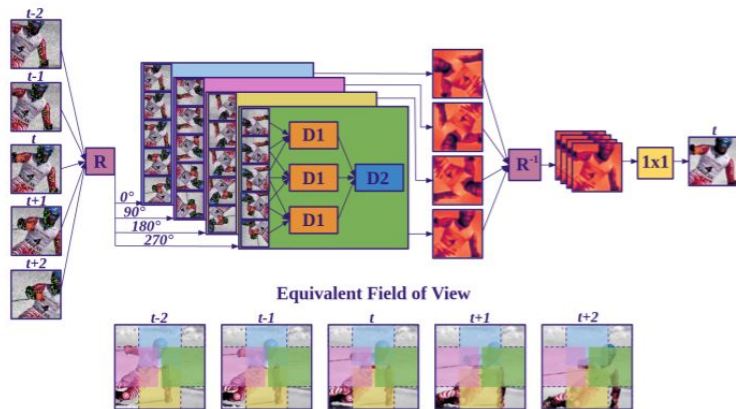


Figure 2. **Unsupervised Deep Video Denoising (UDVD) Network Architecture.** The network takes 5 consecutive noisy frames as input and produces a denoised central frame as output. We rotate the input frames by multiples of 90° and process them in four separate branches with shared parameters, each containing asymmetric convolutional filters that are *vertically causal*. As a result, the branches produce outputs that only depend on the pixels above (0° rotation, blue region), to the left (90° , pink region), below (180° , yellow region) or to the right (270° , green region) of the output pixel. Each branch consists of a cascade of 2 Unet-style blocks (D1 and D2) to combine information over frames. These outputs are then *derotated* and linearly combined (using a 1×1 convolutions) followed by a ReLU nonlinearity to produce the final output. The resulting “field of view” is depicted at the bottom with each color representing the contribution of the corresponding branch.

Multi-frame blind spot architecture

1. Five contiguous noisy frames to a denoised estimate of the middle frame
2. R: rotation
3. D1 :three Unets
 - a. $(t-2)(t-1)(t)$
 - b. $(t-1)(t)(t+1)$
 - c. $(t)(t+1)(t+2)$
4. D2: Unet
5. derotation
6. 1×1 convolution
7. ReLU activation function



Unsupervised Deep Video Denoising

Bias-free architecture



We remove all additive terms from the convolutional layers in UDVD.



This provides automatic generalization to varying noise levels not encountered during training.

Using the missing pixel



The denoised value generated by the proposed architecture at each pixel is computed without using the noisy observation at that location.
=> This avoids overfitting.

In the special case of Gaussian additive noise, we can use this information via a precision-weighted average between the network output and the noisy pixel value.

When the noise process is unknown, we simply minimize the MSE between the denoised output and noisy video, and ignore the center pixel



Unsupervised Deep Video Denoising

In supervised denoising with simulated noise, training can rely on the generation of a virtually unlimited set of fresh noise realizations, which prevents overfitting.

Data augmentation and early stopping



In the unsupervised setting, this is not possible.

- a. leverage data augmentation strategies: spatial flipping and time reversal.
- b. perform early stopping by monitoring the mean squared error between the network output and noisy frames on a held-out set of frames.



These strategies make it possible to train UDVD with short video sequences



Sec4. Datasets



Datasets

We demonstrate the broad applicability of our approach by validating it on domains with different signal and noise structure: natural videos, raw videos, fluorescence microscopy, and electron microscopy

Natural videos

Raw videos

Fluorescence microscopy

Electron microscopy



Sec5. Experiments and Results

Experiments and Results

test set	σ	Traditional		Supervised CNN			Unsupervised CNN (UDVD)		
		VNLB	VBM4D	VNLnet	DVDnet	FastDVDnet	1 frame	3 frames	5 frames
DAVIS	30	33.73	31.65	-	34.08	34.06	32.80	33.48	33.92
	40	32.32	30.05	32.32	32.86	32.80	31.48	32.20	32.68
	50	31.13	28.80	31.43	31.85	31.83	30.47	31.20	31.70
Set8	30	31.74	30.00	-	31.79	31.60	30.91	31.62	32.01
	40	30.39	28.48	30.55	30.55	30.37	29.63	30.42	30.82
	50	29.24	27.33	29.47	29.56	29.42	28.65	29.47	29.89

Table I. **Denoising results on natural video datasets.** All networks are trained on the DAVIS train set. Performance values are PSNR of each trained network averaged over held-out test data. UDVD, operating on 5 frames, outperforms the supervised methods on Set8 and is competitive on the DAVIS test set. Unsupervised denoisers with more temporal frames show a consistent improvement in denoising performance. DVDnet and FastDVDnet are trained using varying noise levels ($\sigma \in [0, 55]$) and VNLnet is trained and evaluated on each specified noise level. All UDVD networks are trained *only* at $\sigma = 30$, showing that they generalize well on unseen noise levels. See Sections C and F in the supplementary material for additional results. The PSNR values for all methods except UDVD are taken from [39].

Experiments and Results

	$\sigma = 30$				$\sigma = 90$			
	DAVIS	Set8	Derfs	Vid3oC	DAVIS	Set8	Derfs	Vid3oC
UDVD-S	33.68 / 78.16	32.90 / 81.85	33.95 / 81.91	34.65 / 84.60	29.05 / 53.53	28.07 / 55.35	29.42 / 59.25	29.94 / 63.79
UDVD*	33.78 / 79.88	31.90 / 82.53	32.58 / 81.44	34.24 / 83.96	28.87 / 51.22	27.25 / 51.84	28.26 / 52.44	29.23 / 60.08
FastDVDnet*	33.91 / 76.99	31.81 / 80.21	32.45 / 81.64	35.05 / 84.44	28.01 / 47.53	26.54 / 50.16	27.36 / 52.87	28.42 / 55.99
MF2F	33.91 / 80.01	31.84 / 80.55	32.87 / 82.22	35.18 / 85.71	28.81 / 51.24	27.25 / 52.78	28.29 / 55.06	29.67 / 61.28

Table 2. **Results for UDVD trained on individual noisy videos.** The top row shows PSNR/VMAF[26] values (averaged over the entire dataset) for UDVD trained on each individual video sequence with early stopping (labelled UDVD-S) using the last 5 frames of a video as a held-out set. We augmented the dataset with spatial flipping and time reversal (see Suppl. D for an ablation study). With the augmentations and early stopping, UDVD-S is comparable to (and often outperforms) UDVD or FastDVDnet trained on the full DAVIS dataset (indicated by *) and MF2F, which fine-tunes a pre-trained CNN on each individual video. See Suppl. D for results on individual video sequences.

Experiments and Results

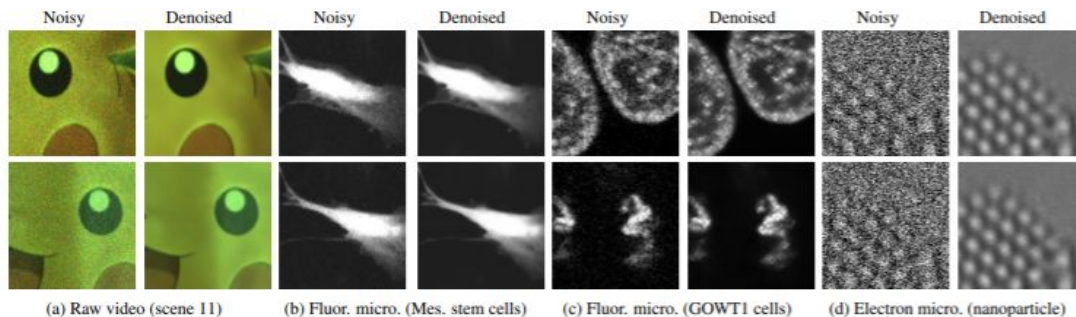



Figure 3. **Denoising real-world data.** Results from applying UDVD to the raw video, fluorescence-microscopy and electron-microscopy datasets described in Section 4. Qualitatively, UDVD succeeds in removing noise while preserving the underlying signal structure, even for the highly noisy electron-microscopy data. Raw videos are converted to RGB for visualization. See Suppl. D and F for denoised videos.

ISO CNN	1600	3200	6400	12800	25600	mean
UDVD	48.04	46.24	44.70	42.19	42.11	44.69
RViDeNet [48]	47.74	45.91	43.85	41.20	41.17	43.97

Table 3. **Raw video denoising.** PSNR values evaluated on the test set of the raw video dataset (Section 4) when denoised with (a) UDVD trained only the noisy test videos and (b) RViDeNet trained with supervision on a large dataset. The columns correspond to different ISO levels, with larger levels resulting in noisier data.

*ISO: International Standardization Organization



Sec6. Automatic Motion Compensation



Automatic Motion Compensation

Gradient-based analysis

A first-order Taylor decomposition of the denoising function may be written as:

$$d_i := f_i(y) = \langle \nabla f_i(y), y \rangle + b,$$



bias-free (i.e., all additive constants are removed from the architecture)

$$d(i) = \langle \nabla f_i(y), y \rangle = \sum_{k=1}^T \langle \underline{a(k, i)}, \underline{y_k} \rangle,$$

equivalent filter / each of the T flattened frames

$\nabla f_i(y) \in \mathbb{R}^{nT}$ denotes the gradient of f_i at y . $b := f_i(y) - \langle \nabla f_i(y), y \rangle$ is the net bias of the network

Automatic Motion Compensation

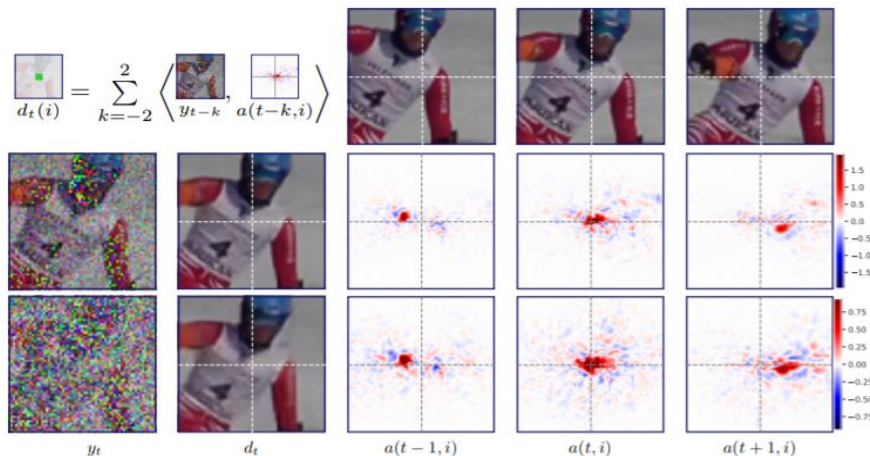


Figure 4. **Video denoising as spatiotemporal adaptive filtering.** Visualization of the equivalent linear weights ($a(k, i)$, Eq. 4) used to compute two example denoised pixels using UDVD. The left two columns show noisy frames y_t at two noise levels, and the corresponding denoised frames, d_t . Three successive clean frames $\{x_{t-1}, x_t, x_{t+1}\}$ are shown in top row, for reference. Corresponding weights $a(k, i)$ for pixel i (intersection of the dashed white lines) in these three frames, are shown in the last three columns. The weights are seen to adapt to underlying video content, with their mode shifting to track the motion of the skier. As the noise level σ increases (bottom row), their spatial extent grows, averaging out more of the noise while respecting object boundaries. For each denoised pixel, the sum of weights (over all pixel locations and frames) is approximately one, and thus can be interpreted as computing a local average (but note that some weights are negative, depicted in blue).

Interpreting equivalent filters

UDVD learns to denoise by performing averaging over an adaptive spatiotemporal neighborhood of each pixel.

Automatic Motion Compensation



Figure 5. CNNs trained for denoising automatically learn to perform motion estimation. (a) Noisy frame from a video in the DAVIS dataset. (b) Optical flow direction at multiple locations of the image obtained using a state-of-the-art algorithm applied to the clean video. (c) Optical flow direction estimated from the shift of the adaptive filter obtained by differentiating the network, which is trained exclusively with noisy videos and no optical flow information. Optical flow estimates are well-matched to those in (b), but deviate according to the aperture problem at oriented features (see black vertical edge of bus door), and in homogeneous regions (see bus roof, top right).

Optical flow estimation

We use the equivalent filters of the networks to estimate the optical flow.

To estimate the optical flow from the (t) th frame to the $(t+1)$ th frame at the i th pixel

↓
Compute the difference between the position of the centroid of the equivalent filter corresponding to the pixel at times t , $a(t,i)$, and time $t+1$, $a(t+1,i)$



Sec7. Conclusion



Conclusion

- In this work we propose a method for unsupervised deep video denoising that achieves comparable performance to state-of-the-art supervised approaches.
- Combined with data-augmentation techniques and early stopping, the method achieves effective denoising even when trained exclusively on short individual noisy sequences, which enables its application to real-world noisy data.
- we perform a gradient-based analysis of denoising CNNs, which reveals that they learn to perform implicit adaptive motion compensation.

Thank you
