Model-blind Video Denoising Via Frame-to-frame Training

Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo and Pablo Arias
thibaud.ehret@ens-cachan.fr
CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France

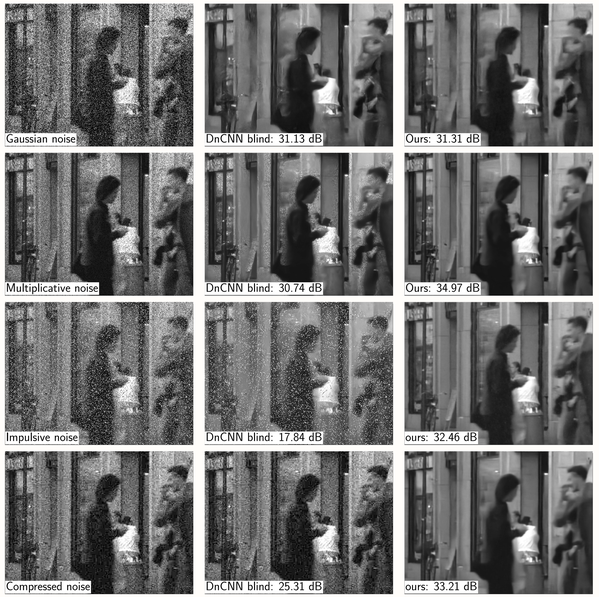
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Contribution

In many practical applications noisy images/videos are the result of a an unknown processing pipeline (demosaicking, gamma, compression, filters). The resulting noise might be difficult to model: spatially correlated, signal dependent. We propose a "frame-to-frame" training strategy that learns how to denoise a video sequence on-the-fly.

Contributions:

- Fine-tuning a pre-trained network to an **unknown noise** model "a la zero-shot": from a **single** noisy video
- Convergence after **only a few frames** (around 50) to **same performance** than a noise-specific network.
- life-long learning: **adapts** to time-varying noise model



Review of Noise2Noise [1]

A **denoising network** can be trained by penalizing the loss between the prediction $\mathcal{F}(f_i)$ and a **second independent noisy observation** g_i of the same image (instead of the ground truth clean image u_i):

$$\mathcal{R}(\mathcal{F}) = \sum_i \{\ell(\mathcal{F}(f_i), g_i)\}$$

Different losses yield different estimators:

$$\begin{aligned} \ell = L_2 \quad &\Rightarrow \quad \mathcal{F}^*(f) = \mathbb{E}\{g|f\} \\ \ell = L_1 \quad &\Rightarrow \quad \mathcal{F}^*(f) = \text{median}\{g|f\} \\ \ell = L_0 \quad &\Rightarrow \quad \mathcal{F}^*(f) \approx \text{mode}\{g|f\} \end{aligned}$$

Under some assumptions on the noise, these estimators coincide with the ones obtained by training with ground truth noiseless data.

Method: online fine-tuning

At every frame, we update the weigths θ by minimizing a **frame-to-frame masked l_1-loss**:

$$\ell_1(\mathcal{F}_\theta(f_t), f^w_{t-1}, \kappa_t) = \sum_x \kappa_t(x) \left|\mathcal{F}_\theta(f_t)(x) - f^w_{t-1}(x)\right|.$$

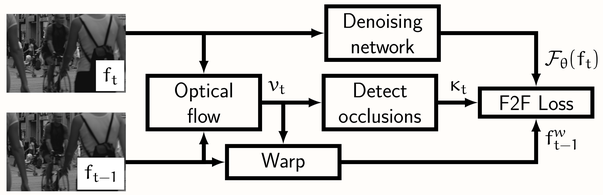
where

- f_t is the **noisy frame** at t
- f^w_{t-1} is the noisy frame at $t-1$ warped to match f_t. The **warping computed using an optical flow** v_t from frame t to $t-1$ (we used TV-L1 [3]).
- $\kappa_t(x) = \mathbf{1}(|\text{div}\, v_t(x)| < \tau)$ are **occlusion masks**

F2F online fine-tuning

for $t = 2, \dots, T$ **do**
  $v^b_t \leftarrow$ `optical-flow`$(f_t, \hat{f}_{t-1}, \sigma)$
  $\hat{f}^w_{t-1}, \kappa_t \leftarrow$ `warp-bilinear`$(\hat{f}_{t-1}, v^b_t)$
  for $i = 1, \dots, N$ **do**
    $\theta \leftarrow$ `adam-step`$(\ell_1(\mathcal{F}_\theta(f_t), f^w_{t-1}, \kappa_t))$
  $o_t \leftarrow \mathcal{F}_\theta(f_t)$

A **batch** fine-tuning is also available.

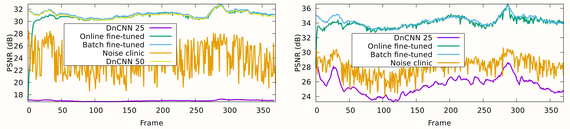


PSNRs curves from synthetic examples

Every fine-tuning started with DnCNN for Gaussian $\sigma = 25$.

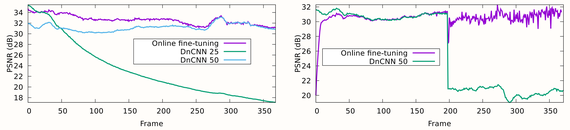
Synthetic noises:

- Left: Gaussian noise with standard deviation $\sigma = 50$.
- Right: Multiplicative noise.



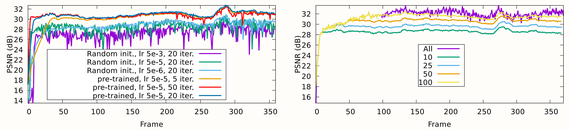
Time-varying noise:

- Left: Slowly evolving noise (Gaussian $\sigma = 25$ to $\sigma = 50$).
- Right: Sudden change (Gaussian σ=50 to impulsive at fr. 200).



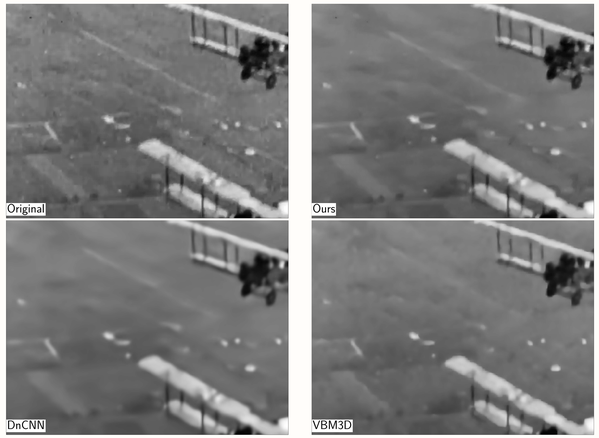
Impact of parameters:

- Left: Impact of pre-training, learning rate and number of iterations.
- Right: Impact of early stopping.

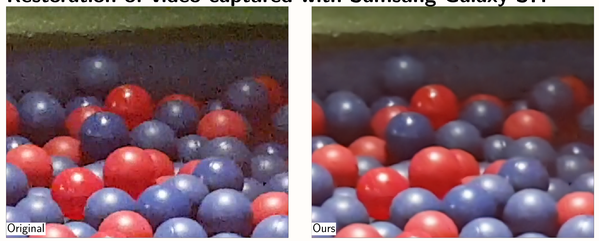


Real examples

Restoration of WW1 footage:



Restoration of video captured with Samsung Galaxy S7:



Code: `https://github.com/tehret/blind-denoising`

References

[1] Lehtinen, Jaakko, et al. "Noise2Noise: Learning Image Restoration without Clean Data.". *ICML*, 2018.
[2] Zhang, Kai, et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising.". *IEEE TIP*, 2017.
[3] Zach, Pock, and Bischof. "A duality based approach for realtime TV-L 1 optical flow.". *Joint pattern recognition symposium*, 2007.
[4] Zamir, Amir R., et al. "Taskonomy: Disentangling task transfer learning.". *CVPR*, 2018.