Self-supervision versus synthetic datasets: which is the lesser evil in the context of video denoising?

Valéry Dewil, Arnaud Barral, Gabriele Facciolo, Pablo Arias

Self-supervision versus synthetic datasets: which is the lesser evil in the context of video denoising?

Valéry Dewil Arnaud Barral Gabriele Facciolo Pablo Arias
Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, 91190, Gif-sur-Yvette, France
https://centreborelli.github.io/VDU2022-the-lesser-evil

Abstract

Supervised training has led to state-of-the-art results in image and video denoising. However, its application to real data is limited since it requires large datasets of noisy-clean pairs that are difficult to obtain. For this reason, networks are often trained on realistic synthetic data. More recently, some self-supervised frameworks have been proposed for training such denoising networks directly on the noisy data without requiring ground truth. On synthetic denoising problems supervised training outperforms self-supervised approaches, however in recent years the gap has become narrower, especially for video. In this paper, we propose a study aiming to determine which is the best approach to train denoising networks for real raw videos: supervision on synthetic realistic data *or* self-supervision on real data*. A complete study with quantitative results in case of natural videos with real motion is impossible since no dataset with clean-noisy pairs exists. We address this issue by considering three independent experiments in which we compare the two frameworks. We found that self-supervision on the real data outperforms supervision on synthetic data, and that in normal illumination conditions the drop in performance is due to the synthetic ground truth generation, not the noise model.*

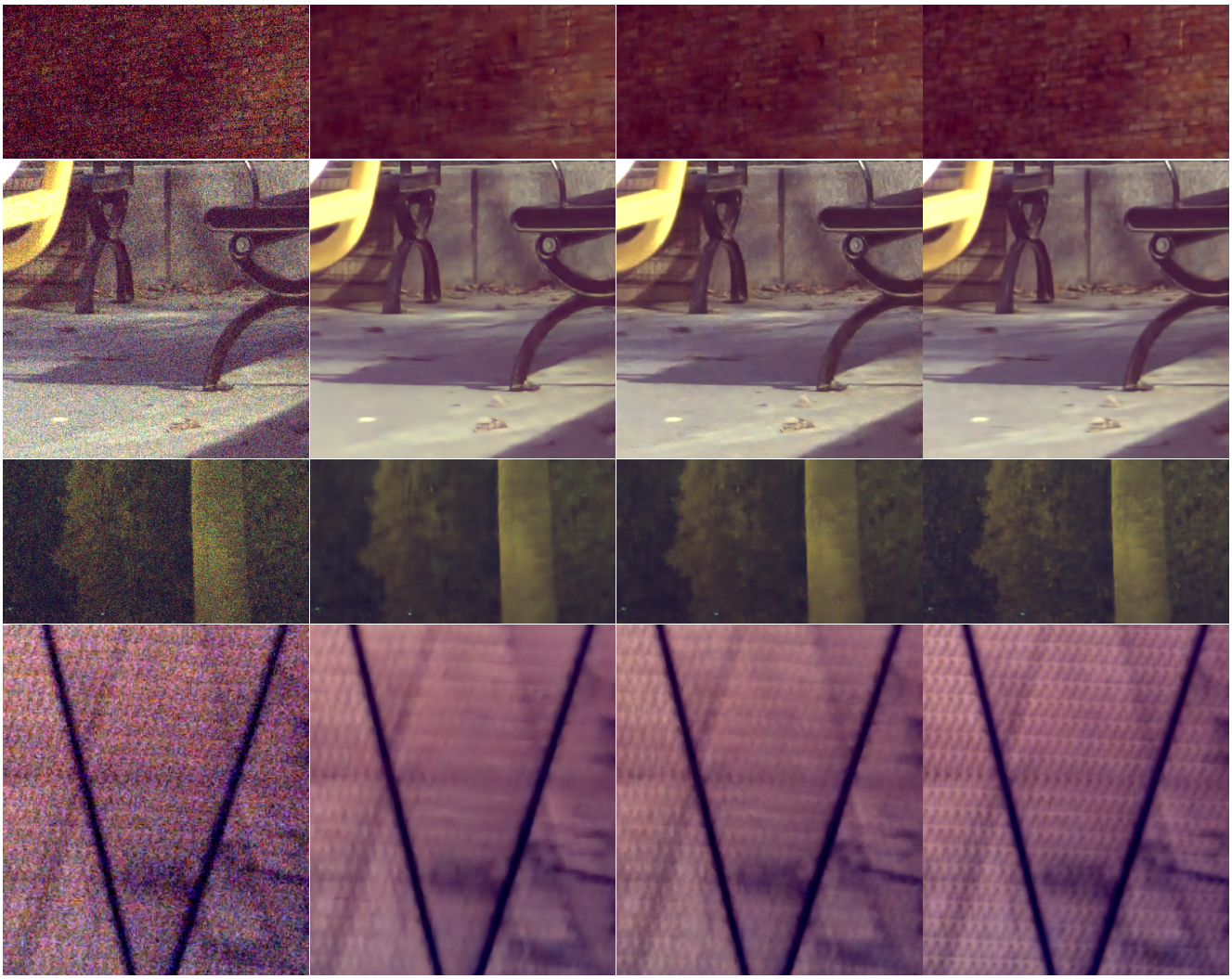
1. Introduction

For both images and videos, denoising is still an active research subject. All the more so in the case of real noise, where the real distribution of the noise may be unknown or at least hard to model. In recent years, data-driven methods based on training convolutional neural networks (CNNs) have taken over the state of the art in several image and video restoration tasks. In addition to their superior performance, CNNs offer a greater flexibility as they can be trained to denoise potentially any type of noise [7,8,22,43]. In contrast, traditional model-based approaches require a tractable model of the noise, and specific algorithms for each type of noise, *e.g.* [4,11,17,28,30,40,52]. This makes learning-based approaches an ideal candidate for restoration of real videos. Yet, this is still rather unexplored, with most of the research that considers real data focusing on single image denoising. The main reason for this is the lack of available training datasets for video denoising.

The standard approach to data-driven methods is via supervised learning, for which a dataset of pairs of input and expected output is used to train the network. Network architectures trained in a supervised manner yield state-of-the-art results. However, supervised training requires large datasets with pairs of clean-noisy signals, which are very hard to obtain in the case of real images [2,8,37], and even more so for real videos. The classical solution is to train networks on synthetic datasets where a clean video is artificially degraded. Nonetheless, CNNs are very sensitive to mismatches between the data distributions at training and testing times [37]. Addressing this issue is currently one of the most important problems in learning-based image restoration, and has recently attracted a lot of attention.

One research trend focuses on generating realistic synthetic datasets for supervised training. The noise in the raw sensor is commonly approximated as an additive heteroscedastic Gaussian with a signal dependent variance [16]. A more accurate model is the Poisson-Gaussian model [16], which still has some limitations as it does not take into account non-linear behavior of the sensor (*e.g.* clipping), dead pixels, heavy tails of read noise, *etc*. It has been shown that more comprehensive models of the noise yield better results [45]. Other works rely on data driven generative approaches to synthesize noise [1,6,9,24,46,48]. Creating synthetic datasets requires synthesizing the clean data too. This is straightforward for RGB denoising, but it is far from trivial for raw denoising [5,44,50] or for other imaging modalities.

Another research trend is based on developing self-supervised approaches that do not require any ground truth, *i.e.* they can be trained directly on the degraded data. An additional advantage is that, in principle, no complex noise modeling is required in order to apply these methods. The



(a) real noisy (b) model-supervised (c) MF2F (d) Self-supervised blind-spot

Figure 1. Comparison of video denoising networks trained with supervision on synthetic data (b) or self-supervision on real data (c-d). All network architectures are based on UDVD [41], MF2F (c) uses the self-supervised framework of [13] and blind-spot (d) uses [41]. *(top-brick wall ISO 3200)* Self-supervised networks recover more details. *(middle top-bench ISO12800)* Natural texture of the stones and the granularity of the ground are removed by the supervised network. *(middle bottom-trees ISO 3200)* Self-supervised networks have a better reconstruction of the texture of the trees *(bottom-wire-grid ISO12800)* The structure of the wire grid is better reconstructed with the self-supervised networks.

main principle of these techniques is to exploit the signal regularity, and train the network to predict one part of the signal from the rest. Self-supervised methods exist for both images [3, 12, 15, 25, 32, 39] and videos [13, 41] denoising, demosaicing [14] and super-resolution [34,35]. On artificial datasets, supervised training outperforms self-supervised approaches. However, recent self-supervised methods have shown competitive results, specially in video denoising.

The natural question is then, *what is the best approach to train denoising neural networks for real videos?* Is it better to train with ground truth supervision on realistic synthetic datasets, or should one train directly on the real data with a self-supervised approach? The former suffers from the generalization gap between simulated and real data, while the latter pays the price of not having supervision from a clean ground truth. The question is which is the lesser evil.

Contribution. In this paper, we study the question of which training framework has to be used for video denoising networks: supervised on synthetic data or self-supervised on real data. This requires to compare quantitatively and fairly both approaches in a controlled setting.

Ideally, this should be done by testing them on evaluation datasets of real natural videos with ground-truth. However, there are no such datasets due to the inherent complexity of simultaneously acquiring noisy and noiseless videos for natural dynamic scenes. We circumvent this problem by considering two surrogates for real data: 1) a synthetic raw dataset with a comprehensive noise model, and 2) a real dataset of static scenes for which ground truth can be estimated via frame averaging. Finally, we evaluate both approaches on real natural videos visually. In all cases, we apply a rigorous training methodology to make sure that we compare fairly the training approaches.

The next section reviews the related work. In Section 3, we present the architecture used in this study as well as a description of the self-supervised trainings. The overall protocol of the study (including datasets and training strategies) is detailed in Section 4. Experiments details and results are presented in Section 5.

2. Related work

Self-supervised training methods are often compared to supervised training on synthetic datasets [3, 13, 15, 25, 41]. In this setting, supervised training is optimal (e.g. with respect to the MSE) and the goal is to achieve the same performance with self-supervision. Our situation is different, since we are interested on the performance on real data of a supervised network trained on synthetic data.

In the case of still images, it is possible to acquire real datasets with ground truth. The ground truth can be either estimated by acquiring a burst of images of static scenes and averaging them [2, 8] or using long exposure times [37]. In such datasets it is possible to train with supervision on real data, and it has been observed that training a network with unrealistic simulated data leads to worse results [1, 45].

This motivated research into how to better simulate real noise. The simplest noise model for raw images is the heteroscedastic Gaussian noise model [29] which supposes the noise to be additive, zero-mean and with a variance as a affine function of the intensity. This corresponds to the sum of two noise sources: the shot noise modeling photons arriving at the sensor and the readout noise introduced by the electronics. In spite of its known limitations, this model is widely used [5, 18, 20, 38, 44, 50]. In [10] the authors use an additive and zero-mean heteroscedastic Gaussian noise but the variance does not follow an affine model. In [51, 53] a Gaussian mixture model is used. In [45], the shot noise is considered Poissonian and a Tukey-Lambda distribution is used to model heavy tails in the readout noise. Additionally, other noise sources are also modeled like the banding pattern noise (e.g. row noise) or quantization noise.

Other approaches for simulating real noisy sequences use data-driven generative methods, such as adversarial generative models [9, 24, 48]. In these works, a generative network is trained to generate a noise close to the real one while a discriminative network is trained to determine whether a noise sample is real or has been generated. In [1] a neural network entirely composed of invertible layers is used to simulate realistic noise from clean data. It was trained on the SIDD dataset [2] and can reproduce the realistic noise of the five cameras with a smaller KL-divergence with respect to the real noise than the heteroscedastic Gaussian noise. Similarly in [46], a CNN is trained to generate realistic degraded data from clean ones.

For raw denoising, it is important also to simulate the raw clean ground truth, a problem that has received less attention. The standard approach is to use sRGB images and apply a simple inverse camera pipeline to generate the raw [5, 18]. In [50] the inverse pipeline is implemented by a network that is learned from real data.

3. Self-supervised training methods

Self-supervised learning methods learn directly from the noisy data by exploiting differences in the correlation structure of signal and noise. A part of the input is withheld from the network, which is trained to predict the withheld part. If the noise of the withheld part is independent from the one given to the network, then the network can only minimize the loss by predicting the clean signal.

The state of the art in self-supervised video denoising is represented by *Multi Frame-to-Frame* (MF2F) [13] and *Unsupervised Deep Video Denoising* (UDVD) [41]. Both techniques were found to be efficient on the real raw video denoising task, and thus we are going to include both of them in our comparison. We will describe them below.

3.1. UDVD: blind-spot network for video denoising

The UDVD method relies on the blind-spot technique recently introduced in [3, 25–27]: a special convolutional network architecture is used, which has a blind-spot at the center of its receptive field. The network is trained to predict the value of this pixel in the noisy video. It is predicted from the surrounding neighbors (both spatial and temporal), exploiting the spatio-temporal regularity of the clean video. The blind-spot technique decreases the denoising performance compared with a non blind-spot (normal) network, as many details are lost. This gap is significant in the case of image denoising, but it is less discernible for video.

Network architecture. For our experiments, we use the architecture of UDVD [41]. This video denoising network takes a stack of five frames as input. It consists of two cascaded Unets (as in [42]). The first Unet is applied three times on each group of three contiguous frames. This produces three outputs that are fused by a second Unet into the final denoised output. The input stack is rotated by the four

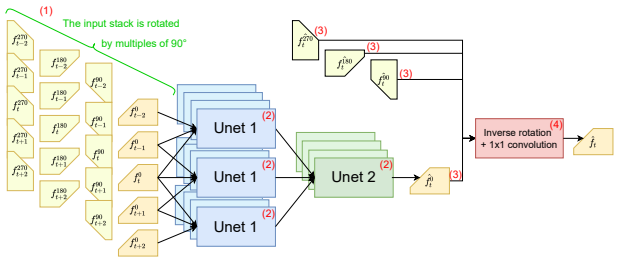


Figure 2. Architecture of the network introduced in [41]. The input stack is firstly rotated by multiples of 90° (1). Each four rotated stacks is processed by the cascaded Unets (2), producing four outputs (3) which are combined together after the rotations are inverted and a 1×1 convolution (4).

multiples of 90° and denoised by the network. The four outputs are finally combined by a 1×1 convolution. The architecture is shown in Figure 2.

To generate the blind-spot, the first Unet uses asymmetric convolutional filters that are vertically causal so that the four outputs only depend on the pixels above. In this way, the receptive field does not contain the central pixel. The blind-spot can be removed by shifting the input data one pixel down after the rotation. The UDVD architecture is also bias-free [31], which has proved to generalize better to unseen noise level at test time.

UDVD training. The self-supervised UDVD blind-spot network is trained by minimizing the L2 loss between the output of the network and the corresponding noisy input frame.

3.2. MF2F training

In MF2F, the weights θ of a network $\mathcal{F}_\theta$ are updated by minimizing the loss $\|\kappa_t \circ (W_{t,t-1}\mathcal{F}_\theta(\mathcal{S}_t) - f_{t-1})\|_1$, where κ_t is an occlusion mask, $W_{t,t-1}$ is a warping operator from frame at time t to $t-1$ (based on an estimated optical flow), $\mathcal{S}_t$ is the stack of frames $[f_{t-4}, f_{t-2}, f_t, f_{t-2}, f_{t+4}]$ and f_{t-1} is the first past frame serving as target (to prevent the trivial identity mapping, it is out of the input stack). The alignment requires a high quality optical flow plus a mask for alignment errors, which are estimated on the noisy data. The MF2F results strongly depends on the optical flow accuracy, which is computed using the TV-L1 method [36,49].

The application of the warping operator $W_{t,t-1}$ requires interpolating the network output at subpixel positions. Interpolating the raw image is problematic. A naive approach would be to pack the raw as a 4 channels image of half the resolution and warp each channel. However, these low resolution channels are heavily aliased. We found better results applying a demosaicing D to the network output, warping on the RGB domain, and re-mosaicing it afterwards. That is, our warping operator can be expressed as $W^{\text{raw}}_{t,t-1} = MW^{\text{rgb}}_{t,t-1}D$, where M is the remosaicing operator. For the demosaicing we use the Hamilton-Adams method [19,21].

4. Methodology

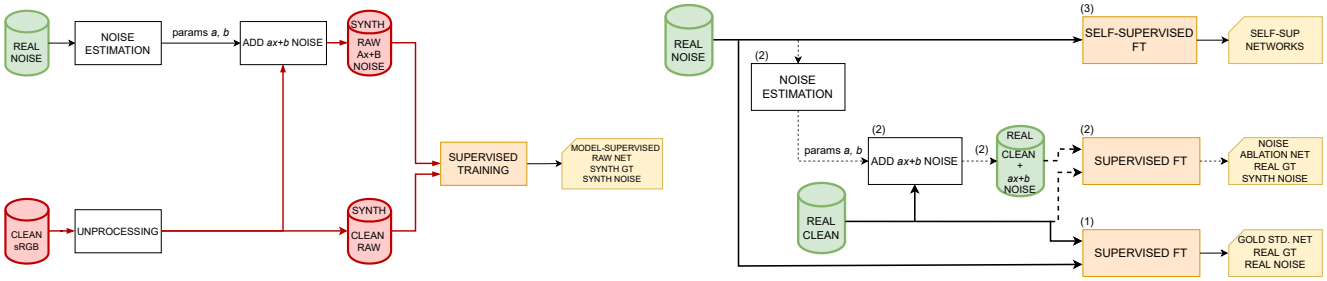
Our goal is to compare two strategies for training a denoising network for raw real videos: supervised training on realistic synthetic data, or self-supervised training directly on the real data. To that aim we need a dataset of synthetic noisy videos and one of real natural videos for evaluation. In the following, we describe our evaluation protocol (see Figure 3).

Dataset of real videos. Since there are no datasets of real natural videos with ground truth, we will consider two surrogates: (1) synthetic videos with a comprehensive noise model, and (2) static real videos with ground truth generated by frame averaging. The first will allow us to measure the effect of an oversimplified noise model in the *synthetic dataset* of dynamic scenes with natural motion. The second is static, but will be useful to have a quantitative evaluation on real data. Lastly, we will consider a dataset of real natural videos for visual evaluation. More details about these datasets will be given respectively in Sections 5.1, 5.2 and 5.3. For simplicity we will talk about the *surrogate real dataset* (abridged to *surrogate dataset*) in the following, even though it might not be actually real data, but our proxy for real data. The *surrogate real dataset* is represented in green in the diagrams of Figure 3.

Dataset of synthetic videos. For each *surrogate dataset* we generate a synthetic realistic dataset with clean-noisy pairs for supervised training (represented in red in Figure 3). We use the REDS 120 dataset [33] which consists of 270 dynamic sequences (split in 240 training and 30 validation sequences) of outdoors scenes taken in daylight conditions, with frame rate 120 fps and of size 1280×720. We temporally downsampled each sequence by taking one frame over three, resulting sequences with 166 frames at 40 fps. Note that this makes the task more complicated for MF2F whose results highly depend on the optical flow estimation accuracy, and therefore on the amount of motion.

These clean sRGB sequences are unprocessed back to the raw domain following [5], which we adapted to our specific case. First, we use a fix color correction matrix throughout all the dataset. We sample random white balance coefficients for each sequence, and use the same coefficients for all frames in the sequence. In the supplementary material, we give more details about our unprocessing procedure steps. This gives us a dataset of clean raw video sequences.

Finally, we add realistic noise to the unprocessed



(a) Pre-training on the *synthetic dataset*

(b) Fine-tuning on the *surrogate dataset*

Figure 3. The *surrogate real dataset* is the green cylinder, the *synthetic dataset* is in red. (a) The model-supervised is trained with supervision on the *synthetic dataset* with synthetic noise (either with or without blind-spot). (b) The previous model-supervised are fine-tuned on the *surrogate dataset*. The steps are (1) fine-tune on real data (2) (when possible) fine-tune on real clean data but with synthetic noise (3) Self-supervised fine-tuning directly on noisy data (UDVD and MF2F).

ground-truth for simulating real noisy sequences from the clean ones. For that purpose, a heteroscedastic Gaussian noise model is estimated from the *surrogate dataset* [29]. More details about the noise estimation can be found in the supplementary material.

Note that the *synthetic dataset* is tailored to model the *surrogate dataset*: we use the same Bayer pattern, the ranges of both datasets are matched and the parameters of the synthetic noise model are fitted to approximate the noise in the *surrogate dataset*.

4.1. Networks

We will use for our experiments the UDVD architecture described in the previous section. This network is computationally costly and has a significant memory footprint. In this paper, we do not focus on achieving the state of the art and reduce this architecture by a factor 4 by using 1/4 of the channels in all layers. This architecture can be used with or without the blind-spot.

Due to the small size of the *surrogate dataset*, we followed [13, 47] and pre-trained the network with supervision on the bigger *synthetic dataset*.

We pre-trained this architecture with a blind-spot as well as without the blind-spot (denoted as *normal*). The reason is that we do not need a blind-spot network for applying MF2F as well as for other supervised training strategies discussed later; while the self-supervised UDVD requires a blind-spot.

For comparing the supervised and self-supervised frameworks, we consider different training strategies. Figure 3 summarizes them. Note that once trained, the evaluation of networks trained with or without supervision requires the same amount of time and computational resources. We now describe the different networks and how we trained them.

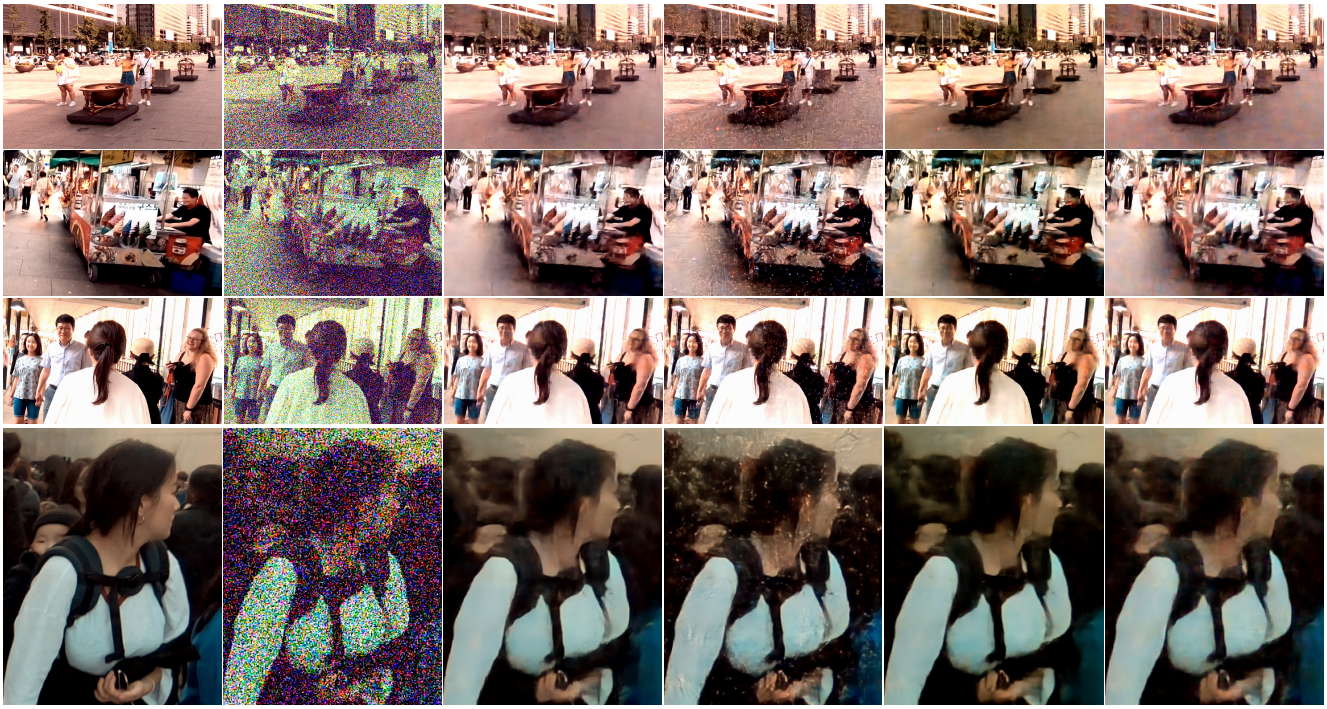
Model-supervised net This network is trained with supervision on the *synthetic dataset*. We train two versions of this network: normal and with the blind-spot. The latter will be used as the pre-trained network for the self-supervised blind-spot fine-tuning, while the former is *the supervised network trained on synthetic data that we wish to compare with the self-supervised approaches.*

Gold standard net The *gold standard* solution for such training is to train with supervision directly on the real data. Although this is not possible in practice because it requires to have access to a large dataset with clean-noisy pairs, it is possible here to fine-tune the *normal model-supervised net* on the *surrogate dataset*. This will give us a reference of the best training that could be achieved to situate the performance of the other trainings.

Noise-ablation net Two kinds of modeling were used for the supervised trainings: the unprocessing of sRGB to generate synthetic raw clean videos and the noise. When possible, we fine-tune an *noise-ablation net* that will allow us to differentiate the impact of the noise model from that of the generation of the clean data by eliminating the unprocessing step. To this aim, we generate noisy images by adding synthetic heteroscedastic Gaussian noise to the clean ground truth of the *surrogate dataset*. We fine-tune the normal model-supervised net on this data with supervision from the real ground truth.

Self-Supervised blind-spot We fine-tune the pre-trained UDVD architecture with blind-spot on the *surrogate dataset* with self-supervision following [41].

MF2F net A second self-supervised network is trained following the MF2F framework as explained in the Section 3.2. Given that MF2F does not requires the network to have a blind-spot, we use the weights of the pre-trained *normal model-supervised net* as starting point of the fine-tuning.



(a) gt (b) real noisy (c) Gold standard (d) Model-sup. normal (e) MF2F (f) Self-sup. blind-spot

Figure 4. Comparison of the different training strategies for the Exp I.

5. Experiments

In this section, we first describe the setting of each of the three experiments together with the obtained results for both approaches. For better visualization, the video frames displayed in this section have been gamma corrected (with $\gamma = 2.2$), demosaicked with [23] and white-balanced.

5.1. Exp I: dynamic scenes with simulated noise

In this experiment we use the REDS 120 dataset to generate the clean raw data for both the *surrogate dataset* and the *synthetic dataset*. The difference lies on the noise model: we use the Poisson-Tukey lambda distribution of [45] as noise model for the *surrogate dataset*. This noise models extreme low-light conditions. In [45] the authors provide parameters for three cameras. We use the noise parameters estimated for the Nikon D850. The noise in the *synthetic dataset* is the heteroscedastic Gaussian with parameters set to approximate the Poisson-Tukey lambda noise of the *surrogate dataset*. All networks are pre-trained on the training split of the *synthetic dataset*, and the fine-tunings are performed using the training split of the *surrogate dataset*.

Results. The first row of Table 1 summarizes the average PSNR on our *surrogate* validation set for the different training strategies. The results show that the self-supervised approaches outperform the supervised training in the *synthetic dataset*: the self-supervised blind-spot network surpasses the model-supervised network by almost 0.7dB and has a much higher SSIM value. The results of the MF2F network have a PSNR similar to the model-supervised, but has a higher SSIM.

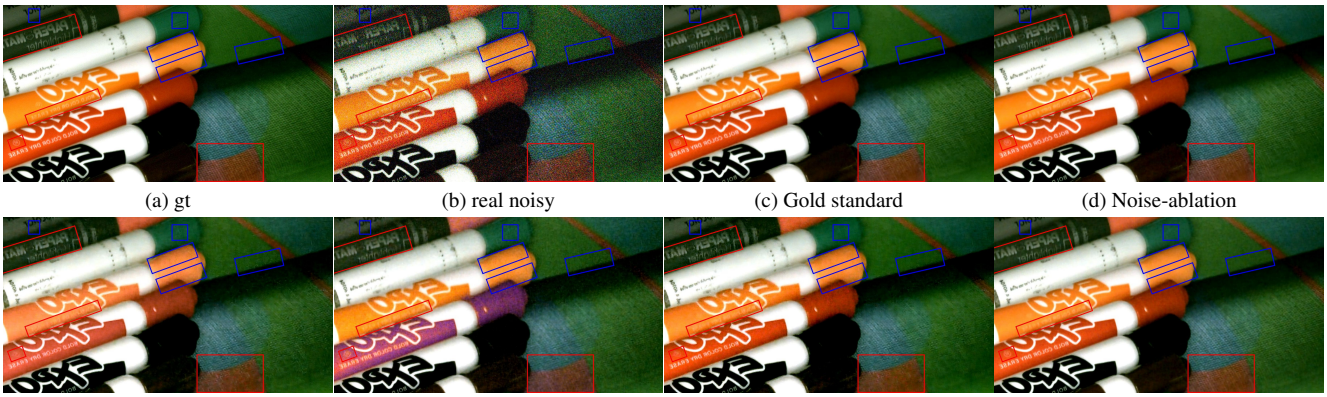
From Figure 4, we notice that both self-supervised networks recover more details and have a better reconstruction of the textures. The self-supervised blind-spot is even close to the gold standard. The result of MF2F has a small color shift, which is why it has a lower PSNR. As the heteroscedastic Gaussian noise model does not fully approximate the noise of the *surrogate real* test set, the model-supervised net results contain denoising artifacts which decrease its performance. On the contrary the self-supervised networks learn the actual noise of this simulated camera and produce results which compete with the gold standard.

5.2. Exp II: real static videos as surrogate data

In the previous experiment, we use artificial ground truth in the *surrogate dataset*. In this section, we are interested in the comparison between model-supervised and self-supervised on real data. To provide quantitative results, we use the *Smartphone Image Denoising Dataset* (SIDD) [2], as it has ground truth. It provides images of ten static scenes, taken by five real cameras with different ISO levels, shutter speeds or illuminations levels. For each,

Exp	Supervised networks				Self-supervised networks	
	Gold standard net	Noise-ablation net	Model-supervised normal	Model-supervised blind-spot	MF2F	blind-spot
I	29.90 / .8393	N/A	28.77 / .7886	27.89 / .8862	28.76 / .8153	29.46 / .9325
II	50.03 / .9913	50.03 / .9913	47.22 / .9847	46.55 / .9938	47.42 / .9849	48.71 / .9965

Table 1. Average PSNR and SSIM over all the sequences of the *surrogate real* test dataset. The model-supervised normal do not have a blind-spot. The model-supervised with blind-spot serves as a pre-trained for the self-supervised UDVD. In Experience I, the *surrogate dataset* is synthetic. Thus we cannot derive an noise-ablation net for this experiment as it would be identical to the model-supervised normal.



(a) gt (b) real noisy (c) Gold standard (d) Noise-ablation

(e) Model-supervised normal (f) Model-supervised blind-spot (g) MF2F (h) self-supervised blind-spot

Figure 5. Comparison of the different training strategies for the Experience II: the normal network trained with supervision (e) recovers less details than networks trained with self-supervision (see the red rectangles) and does not preserve correctly the colors (colors are washed in the orange and red pens). Furthermore, the model-supervised leaves some residual noise (see the blue rectangles).

ISO	Supervised	Self-supervised	
	model-sup. normal	MF2F	blind-spot
3200	42.10 / .9865	43.63 / .9875	41.79 / .9840
12800	35.91 / 9681	38.80 / .9711	37.93 / .9688

Table 2. Evaluation on the indoor CRVD dataset [47]. No network was trained on this dataset (even the self-supervised ones).

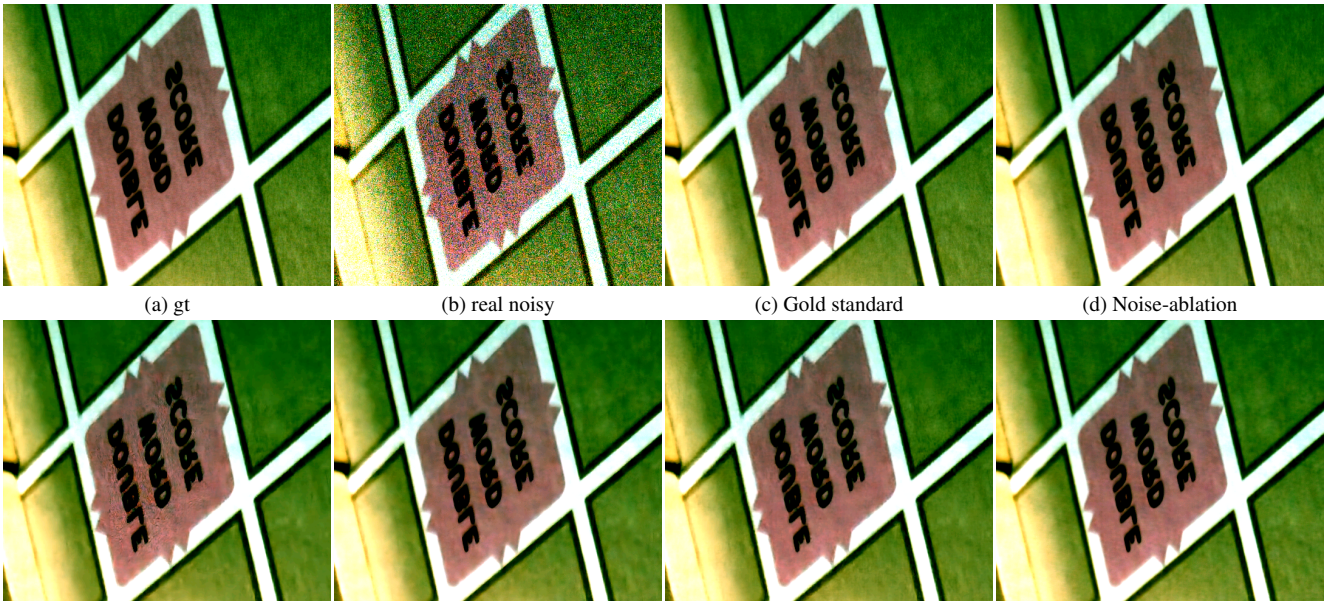
the authors give an estimated ground-truth image obtained by averaging a burst of frames. We generate a ground truth constant video from the ground truth image, as there is no motion in the scene. We use eight static sequences of about 150 frames each obtained with the Google Pixel camera for ISO level 800. This *surrogate dataset* is split into six sequences as a fine-tuning pool and two as a testing/validation pool. For both supervised and self-supervised networks, quantitative results are evaluated on the testing pool of this *surrogate dataset*.

Results. The average PSNR and SSIM on the real validation set are presented in Table 1. As for Experiment I, the trainings with self-supervision lead to a better performance than the trainings done in a supervised setting. On average, the self-supervised blind-spot outperforms the model-supervised by 1.5dB. In Figure 5, the results with the self-supervised networks are sharper and have more details. Figure 6 shows that the model-supervised network creates also artifacts (see near the text).

In this experiment, the *synthetic dataset* differs from the *surrogate* both in the noise model and the ground truth. In order to differentiate this effects, we look at the results of the noise-ablation network, which is trained using clean real data as ground truth, with the simulated heteroscedastic noise. It is remarkable that the result of the noise-ablation network matches exactly with the one of the gold standard (both in PSNR and SSIM). Visual inspection confirms that both results are indeed very similar. We deduce from this that in this case, the heteroscedastic noise model is a good approximation of the real noise, and therefore the problem of the model-supervised network is most likely due to the unprocessed synthetic clean data. On the contrary, in Experiment I, the clean data was the same for both the *surrogate* and *synthetic* datasets, thus the failure of model-supervised net was caused by a bad noise modeling.

5.3. Exp III: real dynamic scenes

In our final experiment, we will use the dataset introduced in [47] for a visual comparison. It consists of *real*



(a) gt (b) real noisy (c) Gold standard (d) Noise-ablation

(e) Model-supervised normal (f) Model-supervised blind-spot (g) MF2F (h) Self-supervised blind-spot

Figure 6. Comparison of the different training strategies for the Experience II: the model-supervised normal leaves some residual noise (see around the letters).

noisy raw videos of 10 outdoor dynamic scenes acquired with a surveillance camera for five ISO levels. For such real data, we do not have ground-truth. As before we pre-train the networks on the *synthetic* REDS 120 dataset with heteroscedastic noise. We considered two ISO levels: 3200 and 12800, and fit the parameters of the noise model to approximate the real noise for each ISO level.

Results. Visual results are shown in Figure 1. In this setting as well, the self-supervised training yields more details leading to a better global reconstruction of the objects.

In [47], the authors also acquire a dataset of videos with ground-truth of indoor scenes taken with the same camera (denoted CRVD). To simulate motion the authors produced stop-motion videos: the camera is fixed on a tripod and several images are taken for ground-truth estimation via averaging. Then, objects in the scene are slightly moved and the process is repeated to acquire new frames. This results in an unnatural motion. As an additional study, we evaluated the previous networks (trained on either the *synthetic dataset* or the *real* outdoor data with real motion) on this indoor dataset. No fine-tunings were done to the indoor dataset as it is very small (10 sequences of only 7 frames each). In particular, the self-supervised networks were trained for the CRVD outdoor dataset and all the networks were trained for real motion. This study is another illustration of the network behavior in case of dataset bias. The quantitative results for two ISO levels 3200 and 12800 are gathered in Table 2. For both ISOs, self-supervised outperforms the supervised network.

6. Conclusion

In this work we propose a protocol to compare in fair conditions two training approaches for denoising real raw videos: supervised training on synthetic data and self-supervised training on the real data. The difficulty of acquiring real videos with ground truth prevents us for doing a simple comparison. To address this issue, we set three experiments covering different use cases such as low light conditions, real motion, real noise at different ISO levels. In all cases, the self-supervised approaches outperformed the supervised one. Among self-supervised techniques, the blind-spot approach UDVD gave better results than MF2F. The main caveat of UDVD is that blind-spot networks tend to be costlier. MF2F can be used to train any multi-frame network architecture. Our experiments also shed light on how to improve the supervised approach. For normal illumination conditions (such as in the SIDD dataset) the main cause of the generalization gap of supervised training on synthetic data, is not necessarily the simple heteroscedastic Gaussian noise, indicating that more effort needs to be put in better modeling of the clean raw data.

Acknowledgments. Work supported by a grant from MENRT. It was also partly financed by Office of Naval research grant N00014-17-1-2552. This work was performed using HPC resources from GENCI–IDRIS (grants 2022-AD011012453R1 and 2022-AD011012458R1) and from the "Mésocentre" computing center of CentraleSupélec and ENS Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.centralesupelec.fr/).

References

[1] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3165–3173, 2019. 1, 3

[2] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 1, 3, 6

[3] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Journal of Machine Learning Research*, pages 524–533, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 2, 3

[4] Jérôme Boulanger, Charles Kervrann, Patrick Bouthemy, Peter Elbau, Jean-Baptiste Sibarita, and Jean Salamero. Patch-based nonlocal functional for denoising fluorescence microscopy image sequences. *IEEE Transactions on Medical Imaging*, 29(2):442–454, 2009. 1

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11036–11045, June 2019. 1, 3, 4

[6] Ke-Chi Chang, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang, and Hwann-Tzong Chen. Learning camera-aware noise models. In *European Conference on Computer Vision*, pages 343–358. Springer, 2020. 1

[7] Yi Chang, Luxin Yan, Meiya Chen, Houzhang Fang, and Sheng Zhong. Two-stage convolutional neural network for medical noise removal via image decomposition. *IEEE Transactions on Instrumentation and Measurement*, pages 1–1, 2019. 1

[8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[9] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155–3164, June 2018. 1, 3

[10] Michele Claus and Jan van Gemert. Videnn: Deep blind video denoising. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, June 2019. 3

[11] Pierrick Coupé, Pierre Hellier, Charles Kervrann, and Christian Barillot. Nonlocal means-based speckle filtering for ultrasound images. *IEEE Transactions on Image Processing*, 18(10):2221–2229, 2009. 1

[12] Emanuele Dalsasso, Inès Meraoumia, Loïc Denis, and Florence Tupin. Exploiting multi-temporal information for improved speckle reduction of sentinel-1 sar images by deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1081–1084. IEEE, 2021. 2

[13] Valéry Dewil, Jérémy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. Self-supervised training for blind multi-frame video denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2724–2734, 2021. 2, 3, 5

[14] Thibaud Ehret, Axel Davy, Pablo Arias, and Gabriele Facciolo. Joint demosaicing and denoising by overfitting of bursts of raw images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[15] Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo, and Pablo Arias. Model-blind video denoising via frame-to-frame training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[16] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 1

[17] Mario Gonzalez, Javier Preciozzi, Pablo Muse, and Andres Almansa. Joint denoising and decompression using cnn regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018. 1

[18] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *arXiv preprint arXiv:1807.04686*, 2018. 3

[19] John Hamilton and James Adams. Adaptive color plane interpolation in single sensor color electronic camera. U.S. Patent 5,629,734, 1997. 4

[20] Ronnachai Jaroensri, Camille Biscarrat, Miika Aittala, and Frédo Durand. Generating training data for denoising real rgb images via camera pipeline simulation. *arXiv preprint arXiv:1904.08825*, 2019. 3

[21] Qiyu Jin, Yu Guo, Jean-Michel Morel, and Gabriele Facciolo. A mathematical analysis and implementation of residual interpolation demosaicking algorithms. *Image Processing On Line*, 11:234–283, 2021. 4

[22] Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Medical physics*, 44(10):e360–e375, 2017. 1

[23] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Minimized-laplacian residual interpolation for color image demosaicking. In *Digital Photography X*, volume 9023, page 90230L. International Society for Optics and Photonics, 2014. 6

[24] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won Jung. Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 3

[25] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2137, 2019. 2, 3

[26] Alexander Krull, Tomas Vicar, and Florian Jug. Probabilistic noise2void: Unsupervised content-aware denoising. Technical report, 2019. 3

[27] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. 2019. 3

[28] Marc Lebrun, Miguel Colom, and Jean-Michel Morel. The noise clinic: a blind image denoising algorithm. *Image Processing On Line*, 5:1–54, 2015. 1

[29] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Practical signal-dependent noise parameter estimation from a single noisy image. *IEEE Transactions on Image Processing*, 23(10):4361–4371, 2014. 3, 5

[30] Matteo Maggioni, Enrique Sánchez-Monge, and Alessandro Foi. Joint removal of random and fixed-pattern noise through spatiotemporal video filtering. *IEEE Transactions on Image Processing*, 23(10):4282–4296, 2014. 1

[31] Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019. 4

[32] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to Denoise from Unpaired Noisy Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12064–12072, June 2020. 2

[33] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[34] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised multi-image super-resolution for push-frame satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1121–1131, June 2021. 2

[35] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised super-resolution for multi-exposure push-frame satellites. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[36] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013. 4

[37] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017. 1, 3

[38] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural information processing systems*, 31, 2018. 3

[39] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2Self With Dropout: Learning Self-Supervised Denoising From Single Image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1890–1898, June 2020. 2

[40] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local pca. *Journal of mathematical imaging and vision*, 48(2):279–294, 2014. 1

[41] Dev Yashpal Sheth, Sreyas Mohan, Joshua Vincent, Ramon Manzorro, Peter A. Crozier, Mitesh M. Khapra, Eero P. Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 2, 3, 4, 5

[42] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1363, June 2020. 3

[43] Puyang Wang, He Zhang, and Vishal M Patel. Sar image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12):1763–1767, 2017. 1

[44] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 1, 3

[45] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3, 6

[46] Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 94–103, 2021. 1, 3

[47] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2301–2310, June 2020. 5, 7, 8

[48] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision*, pages 41–58. Springer, 2020. 1, 3

[49] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 4

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2696–2705, June 2020. 1, 3

[51] Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang. Robust principal component analysis with complex noise. In *International conference on machine learning*, pages 55–63. PMLR, 2014. 3

[52] Weiying Zhao, Charles-Alban Deledalle, Loïc Denis, Henri Maître, Jean-Marie Nicolas, and Florence Tupin.

Ratio-based multitemporal sar images denoising: Rabasar. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3552–3565, 2019. 1

[53] Fengyuan Zhu, Guangyong Chen, and Pheng-Ann Heng. From noise modeling to blind image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2016. 3