Proceedings of the Joint Austrian Computer Vision and Robotics Workshop 2020 DOI: 10.3217/978-3-85125-752-6-33

Real-World Video Restoration using Noise2Noise

Martin Zach, Erich Kobler
Institute of Computer Graphics and Vision
{martin.zach@student, erich.kobler@icg}.tugraz.at

Abstract. *Restoration of real-world analog video is a challenging task due to the presence of very heterogeneous defects. These defects are hard to model, such that creating training data synthetically is infeasible and instead time-consuming manual editing is required. In this work we explore whether reasonable restoration models can be learned from data without explicitly modeling the defects or manual editing. We adopt Noise2Noise techniques, which eliminate the need for ground truth targets by replacing them with corrupted instances. To compensate for temporal mismatches between the frames and ensure meaningful training, we apply motion correction. Our experiments show that video restoration can be learned using only corrupted frames, with performance exceeding that of conventional learning.*

1. Introduction

Recently the approach to signal reconstruction from corrupted measurements shifted from explicitly modeling the statistics of the corruptions and image priors, *e.g*. Block-matching and 3D filtering (BM3D) [6] or Total Variation (TV) based methods [4, 24], to learning based techniques such as Convolutional Neural Networks (CNNs) [11]. Since then, deep learning techniques [9, 18] have become very popular. Residual learning [9], batch normalization [10] and similar improvements along with increasing computational power and high quality datasets made it possible to train such architectures efficiently. Deep architectures are now the state-of-the-art for many image restoration tasks such as denoising, deblurring, and inpainting [8, 13, 19] as well as semantic segmentation [16, 23] and classification [27].

Despite these advances, generalization performance of such models is still largely limited by the size of the available dataset. The acquisition of clean targets is often very tedious or difficult and it has been proposed that data collection is becoming the critical bottleneck in machine learning [22]. It is therefor interesting to investigate whether networks can learn meaningful mappings when only being presented corrupted samples — both as input and as target. Lethinen *et al*. [15] showed that clean targets are not required to learn meaningful reconstructions, provided that the corrupted samples are drawn from an arbitrary distribution conditioned on the clean target which needs to be the expected value. This technique now known as Noise2Noise (N2N) has been successfully applied to image restoration tasks [14].

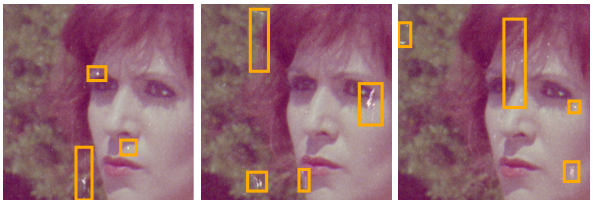

Figure 1. Sample from the dataset, corrupted by typical temporally incoherent and very local defects highlighted in orange.

In this work we explore the applicability of N2N for video denoising, especially concerning the real-world case of having finite data. Due to the nature of the defects, acquiring ground truth samples would require manual editing of the frames and is often not feasible. Further, the defects are very complex and divers in nature such that modeling them is difficult to impossible. Figure 1 displays such an example, where temporally incoherent defects with small spatial extent and high inter-pixel correlation can be seen.

The N2N setting imposes limitations that require special considerations. Since different frames show the scene at different points in time, they cannot directly be used as training pairs. We overcome this by separating temporal motion compensation and spatial denoising, allowing corrupted samples to be both in-

put and target for the model. With this architecture we were able to achieve satisfactory results, showing that video restoration can be done entirely without ground truth data. This significantly eases the task by avoiding the requirement for tedious manual labeling.

2. Related Work

Learning-based Image Restoration Convolutional Neural Networks (CNNs) were first used in 2008 [11], where they achieved similar performance to model based approaches. Later, Burger *et al*. [2] showed that shallow plain Multi Layer Perceptrons (MLP) can achieve results comparable to BM3D. The DnCNN [30] combined recent advances such as the convolutional structure, global residual learning [27], batch normalization [10], and a ReLU activation [20] to achieve a significant performance increase over state-of-the-art explicit models. Later, the FFDNet [31] extended the DnCNN by the use of input noise maps to account for spatially varying noise intensity, in order to apply it to real-world photographs. CBDNet [8] builds on this idea and introduces a noise estimation subnetwork whose output is fed into the denoising network along with the image to achieve notably good results for real-world denoising.

Video Restoration Compared to image denoising, little work exists on video denoising. Patch-based approaches are still the most prominent, *e.g*. V-BM4D [17] and Video Non-Local Bayes (VNLB) [1]. The Deep Video Denoising Network (DVDNet) [28] was one of the first convolutional network approaches to outperform VNLB, whilst being computationally more efficient. In the DVDNet, two separate networks are used for spatial and temporal denoising, and adjacent frames are motion compensated using DeepFlow [29]. Similarly, ViDeNN [5] uses separated spatial and temporal denoising networks, but motion compensation is learned in the temporal network. Frame-to-frame Training [7] exploits N2N by fine-tuning a pretrained network on motion-compensated successive frames. However, the applicability to real-world data remains limited since only one frame is considered for restoration. Besides denoising, learning based methods have been successfully applied to frame interpolation [21], super resolution [3] and deblurring [25].

3. Methods

We consider video scenes $\xi_i = (x_j^i)_{j=1}^{N_f}$ consisting of N_f frames $x_j^i \in \mathbb{R}^{n3}$ with a resolution $n = n_1 \times n_2$ and RGB channels. Each frame of a scene x_j^i is assumed to be corrupted by additive noise, i.e.

$$x_j^i = y_j^i + n_g + n_d \,, \tag{1}$$

where y_j^i is the underlying clean true frame, n_g models noise due to film grain and n_d represents the spatially correlated single-frame defects highlighted in Figure 1. Both noise sources are uncorrelated across the temporal dimension due to the stochastic nature of film grain n_g and the temporal incoherence of n_d. We note that the approach is not limited to this noise model.

3.1. Models for Single-Frame Defect Restoration

The simplest approach to estimate the clean true frame y_j^i is by means of single-frame denoising. For this setting we use the DnCNN [30] to generate a prediction $\hat{y}_j^i$ by

$$\hat{y}_i^j = \mathcal{N}_S^\theta(x_j^i) \,, \tag{2}$$

solely based on the single corresponding corrupted frame x_j^i. Here, θ are the parameters of the DnCNN. They are learned from data either by supervised learning (SL) — provided that target frames are available — or by the N2N approach, which we describe later in this section. The major disadvantage of the single-frame denoising approach is that the model cannot exploit temporal information to detect and restore the single-frame defects.

To overcome this issue and enable the extraction of temporal features, we propose to learn a variant of the DnCNN model operating on two consecutive frames. These two adjacent frames need to be aligned to compensate the motion in dynamic scenes and ease the denoising problem. In detail, we account for the motion by computing the optical flow

$$f_{zj}^i = \mathcal{F}(x_z^i, x_j^i) \tag{3}$$

from frame x_z^i to x_j^i, where $\mathcal{F}\colon \mathbb{R}^{n3} \times \mathbb{R}^{n3} \to \mathbb{R}^{n2}$ implements the pretrained PWC-Net [26]. Using the thereby estimated flow f_{zj}^i, we warp a frame x_z^i of the scene onto the reference frame x_j^i by

$$\hat{x}_{zj}^i = \mathcal{W}(x_z^i, f_{zj}^i) \tag{4}$$

to obtain the motion compensated frame $\hat{x}^i_{zj}$, where $\mathcal{W}\colon \mathbb{R}^{n3} \times \mathbb{R}^{n2} \to \mathbb{R}^{n3}$ is the bilinear warping operator.

In addition, we also compute the backward flow f^i_{jz} and perform a forward-backward check to obtain a binary mask $m^i_{zj} \in \{0,1\}^n$ in the reference frame x^i_j discarding occluded areas. To enable an effective detection of the single-frame defects using temporal information, we require the flow estimation to interpolate over the defects such that they are considered valid in the mask.

Combining the motion compensated frame and the mask with the reference frame x^i_j yields the input to the dynamic model $\mathcal{N}^\theta_D\colon \mathbb{R}^{n3}\times\mathbb{R}^{n3}\times\{0,1\}^n \to \mathbb{R}^{n3}$. Its output

$$\hat{y}^i_{zj} = \mathcal{N}^\theta_D(x^i_j, \hat{x}^i_{zj}, m^i_{zj}) \tag{5}$$

is the estimation of the clean true frame combining spatial and temporal information from two adjacent frames. As before θ denotes the trainable parameters of the DnCNN model learned from data by a SL or N2N approach.

3.2. Supervised and Noise2Noise Learning

Let us first consider supervised learning for reconstructing single-frame defects. Here one requires for every training sample frame x^i_j a corresponding target frame $\bar{y}^i_j$, which can be created by tedious and time-consuming manual editing. Given a collection of corrupted video scenes $\{\xi_i = (x^i_1, \dots, x^i_{N_f})\}_{i=1}^{N_s}$ and a corresponding manually edited target scene $\{\psi_i = (\bar{y}^i_1, \dots, \bar{y}^i_{N_f})\}_{i=1}^{N_s}$, we define the supervised training problem as

$$\min_\theta \sum_{i=1}^{N_s} \mathcal{L}^{\mathrm{SL}}_{\{\mathrm{S,D}\}}(\xi_i, \psi_i, \theta)\,. \tag{6}$$

The scene specific loss $\mathcal{L}^{\mathrm{SL}}_{\{\mathrm{S,D}\}}$ depends on the considered model. For the static model $\mathcal{N}^\theta_S$ we use

$$\mathcal{L}^{\mathrm{SL}}_{\mathrm{S}}(\xi_i, \psi_i, \theta) = \sum_{j=1}^{N_f} \ell\left(\mathcal{N}^\theta_S(x^i_j) - \bar{y}^i_j\right)\,, \tag{7}$$

whereas the loss for the dynamic model $\mathcal{N}^\theta_D$ is given by

$$\mathcal{L}^{\mathrm{SL}}_{\mathrm{D}}(\xi_i, \psi_i, \theta) = \sum_{j=1}^{N_f}\sum_{\substack{z=1\\z\neq j}}^{N_f} \ell\left(\mathcal{N}^\theta_D(x^i_j, \hat{x}^i_{zj}, m^i_{zj}) - \bar{y}^i_j\right)\,, \tag{8}$$

where $\ell \in \{\|\cdot\|_1, \|\cdot\|_2^2, \|\cdot\|_\epsilon\}$ and $\|x\|_\epsilon = \sum_i |x_i|_\epsilon$ is the Huber norm using

$$|x|_\epsilon = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le \epsilon \\ \epsilon(|x| - \frac{1}{2}\epsilon) & \text{else} \end{cases}\,. \tag{9}$$

Despite the constant number of training sample frames, we can use $N_s N_f(N_f-1)$ pairs for training the dynamic model due to the possible permutations, a factor of (N_f-1) more than for the static model.

To avoid the manual editing of target frames, we propose to adopt the N2N approach to remove single-frame defects. Thus, only the corrupted video scenes $\{\xi_i = (x^i_1, \dots, x^i_{N_f})\}_{i=1}^{N_s}$ are used during training. We modify the training problem for N2N to estimate the learnable parameters θ of the models to

$$\min_\theta \sum_{i=1}^{N_s} \mathcal{L}^{\mathrm{N2N}}_{\{\mathrm{S,D}\}}(\xi_i, \theta)\,, \tag{10}$$

using the specific scene loss for the static model

$$\mathcal{L}^{\mathrm{N2N}}_{\mathrm{S}}(\xi_i, \theta) = \sum_{j=1}^{N_f}\sum_{\substack{k=1\\k\neq j}}^{N_f} \ell\left(m^i_{kj} \odot (\mathcal{N}^\theta_S(x^i_j) - \hat{x}^i_{kj})\right) \tag{11}$$

and for the dynamic model

$$\mathcal{L}^{\mathrm{N2N}}_{\mathrm{D}}(\xi_i, \theta) = \sum_{j=1}^{N_f}\sum_{\substack{z=1\\z\neq j}}^{N_f}\sum_{\substack{k=1\\k\neq j\\k\neq j}}^{N_f} \ell\left(m^i_{kj} \odot (\mathcal{N}^\theta_D(x^i_j, \hat{x}^i_{zj}, m^i_{zj}) - \hat{x}^i_{kj})\right)\,. \tag{12}$$

This is illustrated in Figure 2. In contrast to supervised learning, we choose a frame x^i_k and compensate for the motion to the reference frame x^i_j and get the warped frame $\hat{x}^i_{kj}$ as well as the binary mask m^i_{kj}. Then we only evaluate the loss function in the areas where the forward-backward check is consistent to disregard motion estimation errors. A particular advantage of N2N learning is that a factor of (N_f-1) more training samples are available for the static model and (N_f-2) for the dynamic model without the necessity to manually edit any frame.

In all our numerical experiments we optimize (6) and (10) using a dataset of $N_s = 368$ video sequences of $N_f = 3$ frames, which was divided into training (343) and test set (25). For each of the 368 samples there is 1 manually edited target at $j = 2$, where only the single-frame defects n_d were removed and the film

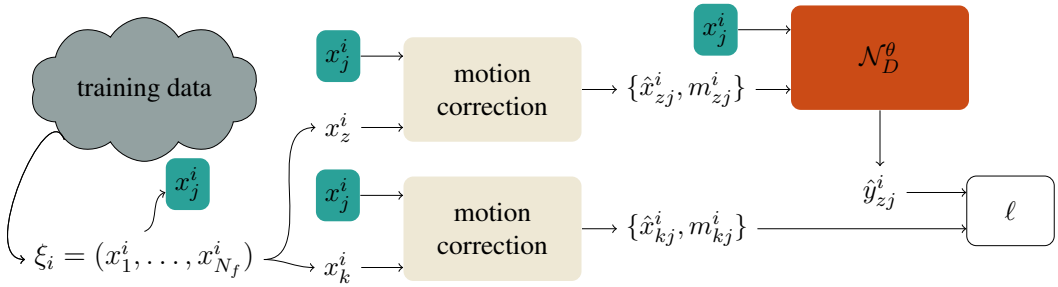

Figure 2. Illustration of the proposed sampling process for N2N learning to video restoration using motion compensation. Here we choose x_j^i as the reference frame, and warp x_z^i and x_k^i onto it. Then, we calculate the estimate $\hat{y}_{zj}^i$ by using the reference frame x_j^i and $\hat{x}_{zj}^i$, and finally the loss using $\hat{x}_{kj}^i$.

Error	static		dynamic	
	SL	N2N	SL	N2N
ℓ_2	0.002 151	0.018 161	0.000 675	0.001 648
ℓ_1	0.002 736	0.012 005	0.000 320	0.001 910
$\epsilon = 0.1$	—	—	0.000 721	0.001 630

Table 1. Evaluation of the average mean squared error to the manually edited target images of the test set.

	Original	SL	N2N
Overall Best	3.13 %	43.23 %	53.65 %
Least Flickering	0.52 %	10.94 %	88.54 %
Significant Smoothing	0 %	1.04 %	56.77 %

Table 2. Quantitative evaluation of the reader study. The results of indicate that the majority of participants prefers the N2N method, where artifacts are significantly better removed at the cost of introducing some smoothing.

grain was not changed. We used a pre-trained PWC-Net [26] for motion compensation and extended the DnCNN [30] to 20 layers with batch normalization, and 64 convolution kernels of size 3×3. Using the ADAM [12] optimizer on a batch size of 128, we trained the models for 3000 iterations with a learning rate of $\alpha = 1 \times 10^{-4}$ and decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We sampled patches of size 64×64 from the frames and augmented the data by vertical and horizontal flipping. Finally, we estimate $\hat{y}_2^i$ as

$$\hat{y}_2^i = \begin{cases} \hat{y}_{12}^i & \text{if } m_{12}^i \wedge (\neg m_{32}^i) \\ \hat{y}_{32}^i & \text{if } m_{32}^i \wedge (\neg m_{12}^i) \\ \frac{\hat{y}_{12}^i + \hat{y}_{32}^i}{2} & \text{else} \end{cases} . \tag{13}$$

4. Results

In this section we present results to highlight the benefits of N2N learning for removing single-frame defects in scanned historical video scenes. We perform quantitative and qualitative evaluation for the static and dynamic models and compare supervised learning to N2N. The qualitative results were also evaluated in a reader study with a focus on temporal coherence.

We show the Mean Squared Error (MSE) on the test set in Table 1 and some representative examples in Figure 3. Given the nature of the defects, their detection is easier if the model can use temporal information. This is confirmed by the results in Table 1, since the results show that the dynamic model outperforms the static model.

The numerical results indicate better performance for the models trained on SL targets. However, this is misleading since it does not necessarily correspond to better defect removal. In fact, Figure 3 suggests that N2N learning improves defect removal. The superior MSE of supervised models is explained by the preservation of film grain, which has not been removed in the targets. In contrast, since film grain differs between the frames, N2N models learn to remove it. Thus, even though they are qualitatively better at removing defects, they yield worse numerical errors.

Further, visual quality of videos cannot be determined by considering the individual frames only. The temporal context needs to be considered as well, where incoherencies can lead to an unpleasant viewing experience. Quality measures could be improved by taking temporal coherency into account, however objective evaluation would still be problematic. Thus, numerical error measures are not suited to fully determine the visual quality of the output.

In general, evaluation is best done by a human who can subjectively decide whether, *e.g.*, removal of film grain is desired, and how pleasant the final video is to watch over all. We therefor conducted a reader study[1]

[1]Material available at https://github.com/zacmar/restoration-reader-study

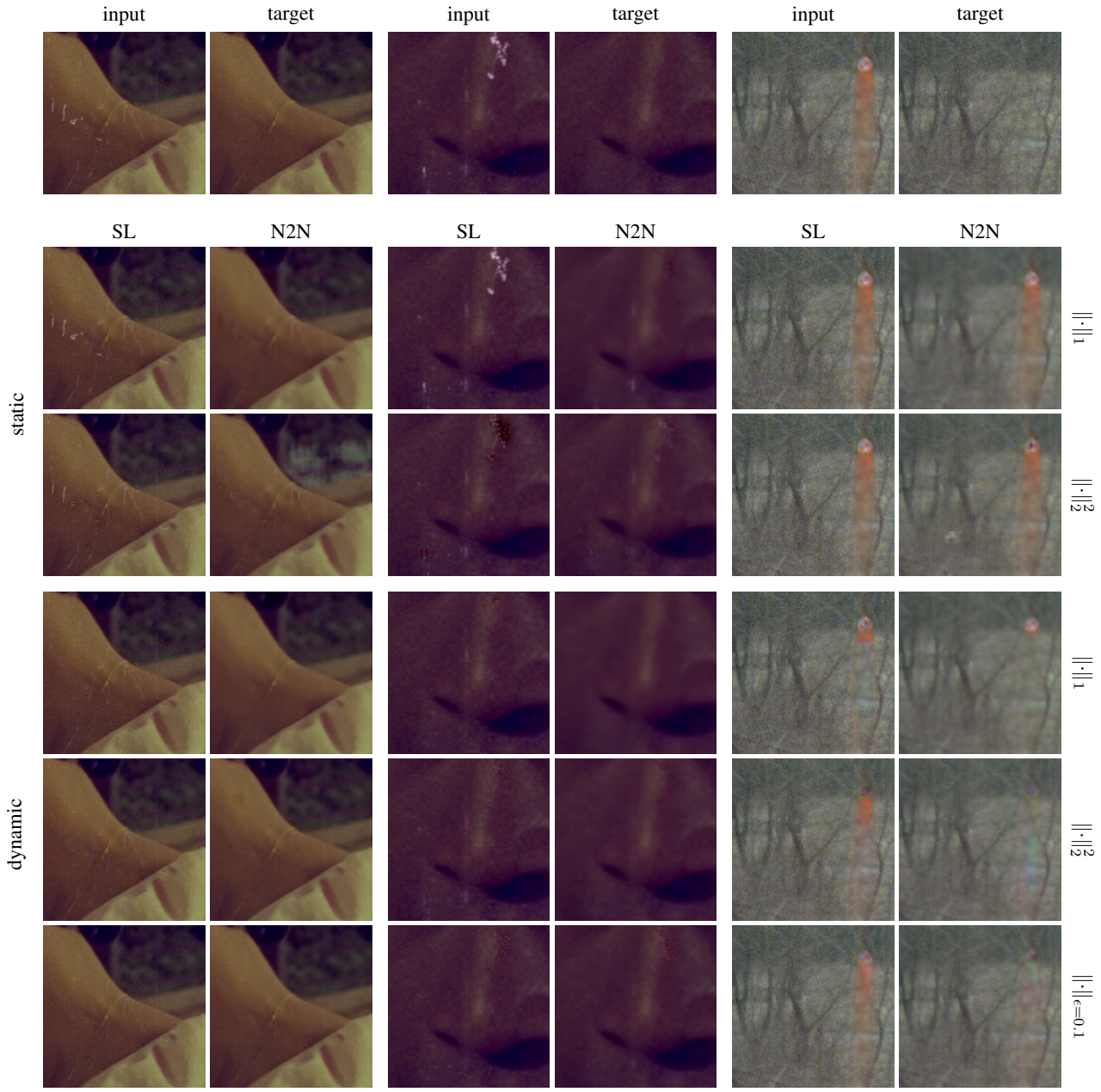

Figure 3. The first row depicts crops from the corrupted frame x_j^i along with the corresponding manually edited target $\bar{y}_j^i$. The second and third row show the results obtained using the static model $\mathcal{N}_S^\theta$, whereas, the results of the dynamic model are depicted in the last three rows. The columns alternate between supervised learning (SL) and N2N results and on the right we show which loss function was used during training.

in which the reader was presented three versions of the same scene side by side: (i) The original frames, the output of the models trained using (ii) SL and (iii) N2N ($\|\cdot\|_\epsilon, \epsilon = 0.1$). Table 2 presents the results obtained from 24 people who were each shown 8 video sequences. It shows that the model trained with N2N is best at removing the defects, at the cost of over smoothing the images. Still, it was the overall preferred method, with $53.65\,\%$ of all samples being deemed "Overall Best" by the participants.

5. Conclusion

In this work we explored the possibilities of using N2N learning for video restoration. We trained static and dynamical models by considering adjacent frames using supervised learning and N2N, relying on robust motion estimation. Using this paradigm we demonstrated that video restoration can be learned by only looking at corrupted frames at performance levels exceeding those of supervised learning. This opens

up new possibilities in areas where acquiring clean training data is too time consuming or infeasible.

There are some limitations that we leave for future research. Due to the structure of our dataset, the number of samples available for N2N learning was limited by the available ground truth targets. Since N2N does not require manual frame editing, it is possible to increase the size of the dataset without much effort. Along with the increase of the size of the dataset, the model complexity could be increased, typically resulting in better performance.

Acknowledgements

The authors acknowledge grant support from the National Institutes of Health under grant 1R01EB024532-02.

References

[1] P. Arias and J. Morel. Video denoising via empirical bayesian estimation of space-time patches. *JMIV*, 60:70–93, 2018.

[2] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with BM3D? In *CVPR*, 2012.

[3] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. *CVPR*, 2017.

[4] A. Chambolle. An algorithm for total variation minimization and applications. *JMIV*, 20(1):89–97, 2004.

[5] M. Claus and J. C. van Gemert. ViDeNN: Deep blind video denoising. In *CVPR Workshops*, 2019.

[6] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IP*, 16(8):2080–2095, 2007.

[7] T. Ehret, A. Davy, J. Morel, G. Facciolo, and P. Arias. Model-blind video denoising via frame-to-frame training. In *CVPR*, 2019.

[8] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[11] V. Jain and H. Seung. Natural image denoising with convolutional networks. In *NIPS*, 2008.

[12] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[13] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock. Variational networks: Connecting variational methods and deep learning. In *GCPR*, 2017.

[14] S. Laine, J. Lehtinen, and T. Aila. Improved self-supervised deep image denoising. In *ICLR*, 2019.

[15] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning image restoration without clean data. In *ICML*, 2018.

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[17] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IP*, 21(9):3952–3966, 2012.

[18] X.-J. Mao, C. Shen, and Y.-B. Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *ArXiv*, abs/1606.08921, 2016.

[19] X.-J. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, 2016.

[20] V. Nair and G. Hinton. Relus improve restricted boltzmann machines. In *ICML*, 2010.

[21] S. Niklaus and F. Liu. Context-aware synthesis for video frame interpolation. In *CVPR*, 2018.

[22] Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *KDE*, 2019.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[24] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1):259 – 268, 1992.

[25] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. *CVPR*, 2017.

[26] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[28] M. Tassano, J. Delon, and T. Veit. DVDnet: A fast Network for Deep Video Denoising. In *ICIP*, 2019.

[29] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. *ICCV*, 2013.

[30] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IP*, 26(7):3142–3155, 2017.

[31] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for CNN based image denoising. *IP*, 2018.