State of the Art on: Deep Image Denoising

Edoardo Peretti, edoardo1.peretti@mail.polimi.it

1. Introduction to the research topic

Recent technological and methodological advances have allowed the employment of deep learning techniques, in particular deep artificial neural networks, in a large variety of fields. One of the fields that most is benefiting from the introduction of deep learning is image processing and computer vision, which mainly exploits convolutional neural networks (CNNs) for addressing visual understanding problems. For instance, the use of CNNs for image classification and object detection has led to outstanding results. In the last years, deep learning models have been successfully employed also for the tasks of image restoration. Starting from a corrupted image (e.g. noisy, blurred), the goal of image restoration is to recover the original image (i.e. the *clean* image). Depending on the type of corruption, image restoration tasks can be divided into deblurring, super-resolution, denoising, inpainting, text removal and many others. In particular, the main focus of our research work is image denoising, which is perhaps the most widely addressed problem in the literature. The major journals and conferences regarding deep learning for image denoising, sorted for h5-index, are the following:

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), h5-index: 240
- Neural Information Processing Systems (NIPS), h5-index: 169
- The European Conference on Computer Vision (ECCV), h5-index: 137
- International Conference on Machine Learning (ICML), h5-index: 135
- International Conference on Computer Vision (ICCV), h5-index: 129
- IEEE Transactions on Image Processing, h5-index: 105

1.1. Preliminaries

The typical representation of an image is by means of a 3-dimensional array of numbers: two dimensions for height and width, and one dimension for the color channels. In principle, images can be flattened and processed with a standard fully connected neural network. However, since images are very high-dimensional data, the number of parameters would soon be extremely large and the network would become untrainable in practice. This is also why images are typically processed with convolutional neural networks. These networks perform, for each pixel, the convolution operation against the same learnable kernels, allowing weights sharing between pixels and reducing the total number of parameters. Each pixel of the output (or of an intermediate layer) will depend only on a portion of the input image. This portion is called *receptive field* and it should be large enough to capture the necessary information for successfully addressing the classification/regression task. Interestingly, fully convolutional networks (which represent the most popular restoration models) can deal with images of arbitrary size in a seamless way, without requiring cropping or resizing. In [7] an extensive presentation of deep learning architectures is provided.

While tasks such as image classification have always been developed with machine learning techniques, image restoration algorithms used to be designed in an expert driven fashion, by means of signal processing techniques and mathematical/statistical modeling of the input image. Only recently, thanks to the power of CNNs, image restoration has been addressed with machine learning techniques.

A variety of libraries are available for developing deep learning models. The earliest works on deep image denoising used to be implemented with the MatConvNet [25] toolbox for Matlab. More recently, Python has

become the main language for implementing artificial neural networks, thanks to the availability of several high-level libraries such as TensorFlow [1] and PyTorch [20]. Typically, GPUs are used to dramatically reduce the training time. Finally, as for each machine learning method, a large amount of training and testing data is required. This latter aspect, as long as synthetic noise removal is concerned, is not a relevant problem because virtually every image dataset can be employed when the noise is added synthetically (one of the most used dataset is BSD [19]). For realistic noise reduction, special datasets, which are expensive to collect, are required (e.g. DND [21]).

1.2. Research topic

A noisy image $\mathbf{y}$ can be modeled as

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\eta}$$

where $\mathbf{x}$ is the latent clean image and $\boldsymbol{\eta}$ is the noise (or *residual*) corrupting the image. The goal of image denoising is to provide an estimate $\hat{\mathbf{x}}$ as close as possible to the clean image, with respect to some metric. The simplest model for $\boldsymbol{\eta}$ is the additive white Gaussian noise (AWGN), where each pixel is corrupted by independent realizations of a fixed Gaussian distribution, that is $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where the standard deviation σ (also known as *noise level*) specifies how strong the noise is. A more realistic signal-dependent noise for raw images is obtained modeling the photon sensing with a Poisson and other stationary disturbances with a Gaussian [6]. One of the main causes of noise is undoubtedly the in-camera processing, for instance demosaicing, gamma correction and compression. This processing makes the noise spatially and chromatically correlated and thus harder to remove. Most of the CNN denoisers are discriminative methods which aim to learn an estimate of the function

$$\mathcal{F}(\mathbf{y}) = \mathbf{x} \tag{1}$$

in order to directly predict the latent image, or of the function

$$\mathcal{R}(\mathbf{y}) = \boldsymbol{\eta}$$

to estimate the noise. Methods that follow the latter approach are said to perform *residual learning*.

The obvious application for image denoising is to provide to the user a pleasant and clear image by removing as much noise as possible, without losing details. From a technical point of view, a denoised image is an essential prerequisite for more high-level computer vision tasks and complex deep learning pipelines (e.g. autonomous driving). Moreover, denoiser can be plugged in as a modular part of model-based optimization methods to solve other restoration problems [31].

2. Main related works

2.1. Classification of the main related works

Denoising methods can be classified from several points of view. Probably, the most immediate classification is between classic methods and methods based on neural networks. The former use priors such as sparsity and self-similarity that have been proved successful over the last few decades in denoising. These priors are typically leveraged thanks to transformation (e.g. DCT, wavelets, etc.) and similarity search procedures such as block matching for patch-wise processing. The latter have recently emerged, exploiting the considerable advances achieved in CNNs training (like ReLU [13], batch normalization [10], residual learning [9] and dilated convolution [28]) and the availability of parallel computing platforms (e.g. GPUs). Machine-learning based methods have recently outperformed classic methods. Nowadays, the research is entirely focusing on learned models and machine learning techniques, which still show room for improvement.

Another possible classification can be done considering the type of noise that the methods aim to remove from the corrupted images. Most of the methods are designed to handle additive white noises (e.g. Gaussian, Poisson, Bernoulli, Salt & Pepper). Few techniques are able to properly remove more realistic noise types which are spatial

and inter-channels correlated, heteroscedastic and signal dependent. Both approaches are still under consideration by the research community, with increasing attention to more realistic noises.

A further classification dimension can be obtained considering whether the denoiser must receive (or assume) some information about the noise distribution of the corrupted image to restore. If this is the case, the denoiser is said to be *non-blind*, otherwise is called *blind*. In case of AWGN, typical information that a non-blind denoiser requires is the standard deviation of the noise (i.e. the *noise level*), but most of the noise distributions can be parametrized by a few parameters.

2.2. Brief description of the main related works

2.2.1 Classic methods

One of the simplest denoising algorithms consists in averaging the pixels by means of convolution with a gaussian kernel. On the one hand this operation suppresses noise in the smooth regions of the image, but on the other hand causes undesired blurring at edges and fine details where neighbouring pixels are not conveying the same information. The most effective classic methods perform transformations over groups of pixels (or rather, groups of patches) that may not be close to each other. Indeed, the non-local means [2] performs a weighted average of all pixels in the image. The famous BM3D [5] algorithm first performs a block matching to group similar patches and then applies a collaborative filtering to remove the noise. Classic methods have achieved good performance in denoising, but they can require manual setting of parameters and their inference is computational expensive. A survey can be found at [11].

2.2.2 Learning based methods

Deep learning methods require a large amount of training images and a computationally very expensive training procedure. On the other hand, their inference time is considerably shorter than those of the classic methods and the denoising performance have greatly surpassed those of classic methods. When the direct mapping $\mathcal{F}$ (1) is similar to an identity mapping, the residual mapping will be easier to learn [30]. This is exactly the case of image denoising because the noisy observation is similar to the clean image and therefore the residual learning formulation is more suitable for image denoising (indeed most of the deep learning methods are residual learners). In the following, a brief overview of some of the more relevant deep learning techniques for image denoising is given. For a more comprehensive and detailed overview refer to [24].

Basic networks IRCNN [31] and DnCNN [30] are simple CNN denoisers, designed mainly for additive white gaussian noise (AWGN), whose architecture is basically a stack of convolutional layers with ReLU activations and zero padding to maintain the spatial dimensions. Both the CNNs introduce batch normalization between the convolutional layers which has been proved to be particularly effective when dealing with AWGN. FFDNet [32] is an extension of DnCNN which operates on images with reduced spatial size (without loss of information thanks to the increased number of channels) in order to speed up the CNN and to enlarge the receptive field. Moreover, FFDNet also expects as input a noise level map that should provide the noise standard deviation for each pixel of the noisy image (thus it can also handle spatial varying Gaussian noises). Thanks to this noise map, FFDNet is a non-blind denoiser and its learned parameters do not depend on the noise levels seen during training. Obviously, if the provided noise level is different from the real one, the inferences are suboptimal: if it is smaller, much noise will remain; on the opposite, image details are smoothed out along with the noise.

More recent solutions The authors of MemNet [23] introduce a concept of memory inside the CNN. They propose a memory block consisting of a recursive unit and a gate unit. The recursive unit is in turn composed of a series of residual building blocks implemented with two convolution layers each. The output of each residual block is concatenated (forming the short-term memory) and forwarded to the gate unit which also receives the concatenation of the outputs of the previous memory blocks (forming the long-term memory). The gate

unit mixes these features (which are multi-level representations under different receptive fields) by means of a 1×1 convolution. They show that dense connections from previous layers can compensate for the loss of mid/high-frequency information in very deep networks.

In [14], the authors show that, from a theoretical point of view, it is possible to train a CNN to denoise images without providing examples of clean images. More in detail, what the CNN learns does not change if the input-conditioned target distributions $p(\mathbf{y}|\mathbf{x})$ are replaced with arbitrary distributions that have the same conditional expected values. This means that we can corrupt the targets with zero-mean noise without affecting the learning. Their experiments show that, in practice, it is possible to achieve comparable results as if clean examples were provided for training. This approach removes the difficulties of collecting clean-noisy pairs for training but requires a large amount of different noisy realization of the same image.

The authors of [22], propose a network that gradually provides intermediate noise estimates which are then all summed up to the noisy images to recover the clean image. More importantly, they trained several models specialized for image contents (e.g. faces, cars, pets, etc). These class-aware denoisers are more effective because they narrow down the space of images to a more specific class and hence they have been trained to restore specific patterns (e.g. the eyes in a face). They also show that using a deep classifier, to choose which denoiser to employ, instead of relying on an oracle, does not deteriorate significantly the performance.

Iterative denoising Some works try to improve the denoising performance in an iterative way. In NN3D [4] a cascade of a CNN and a non-local filter (NLF) is applied iteratively to remove the noise. The NLF complements the weights sharing property of the CNN with nonlocal self-similarity exploitation. They show that this filter cascade can boost the performance of CNNs considered by themselves, especially for images with structures (i.e. having strong self-similarity). Analogously, a deep boosting framework [3] is utilized to improve denoising performance using a CNN as boosting unit. Theoretically, every CNN denoiser can be employed as a boosting unit, in practice special attention must be taken in order to be able to effectively train the model (stacking several boosting units lead to a very deep network which is difficult to train). To this aim, they optimize a CNN to be employed in this boosting framework. In NBreaker [15], the problem of unknown and complex noise is addressed, where the noise is assumed to be a sequential composition of different primary distributions. A CNN classifier is employed to identify the relevant noise and select the corresponding specialized denoiser. This process is repeated until the classifier identifies the "clean" class, namely that there is no noise left in the image. NBreaker does not obtain great results because it strongly suffers from misclassified noises and makes strong assumptions on the noise.

Non-local operations The authors of NLRN [16] propose a differentiable non-local module that can be integrated into existing CNNs to capture feature correlation between each location and its neighborhood. They further propose a recurrent neural network (RNN) for image denoising where the function that updates the recurrent state is implemented with their non-local module and other convolutions. More in detail, they initialized the state by means of a convolutional layer and they fixed a number T of time steps during which the state is updated. The output inference is provided only after the T-th step. Experiments demonstrate that NLRN achieve promising performance, particularly for images with strong self-similarity. In [12] a pixel adaptive filtering unit is proposed. For each pixel, the unit performs a differentiable selection of kernels from a discrete, learnable and decorrelated group of kernels in order to make convolutions content-aware. This unit is used to replace the input convolutional layer of existing CNNs for joint demosaicing and denoising, achieving better results than the corresponding standard CNNs.

Realistic noises The previous methods mainly deal with additive white gaussian noise or other simple noise distributions. Very recently, some deep learning techniques were proposed to deal with more realistic noise distributions. In CBDNet [8], the authors propose a more complex noise model considering the Poisson-Gaussian model [6] and the noise due to in-camera processing (such as demosaicing, gamma correction and compression). With this noise model, they trained a CNN consisting of a noise estimation subnet and a non-blind denoiser, achieving excellent performance on real image denoising benchmarks. In VDN [29], a generative Bayesian

framework is presented, considering the target image noisy itself and introducing a latent clean image. The posterior distributions are then approximated with a variational distribution whose hyper-parameters are estimated with a CBDNet-like CNN. VDN achieves state-of-the-art results in both AWGN removal and in real image denoising.

Conclusion Recent developments have been driven by practical considerations (e.g. training without clean images and more realistic noise models) or by transferring the foundations of the classic methods into data-driven deep models (e.g. non-local filters). It is noteworthy the fact that most of the previous networks can be employed, with small modification if any, to address different noise distributions and also other restoration tasks such as single image super-resolution and JPEG deblocking, provided that the training was performed accordingly.

2.2.3 Metrics and losses for image quality

Regardless of the method employed, metrics are necessary to evaluate and compare the denoising performances. The typical metric is the peak signal-to-noise ratio (PSNR) which is used to compare a degraded image with its clean version. The PSNR is based on the mean square error and therefore it is linked to the ℓ_2 error function. However, these metrics do not correlate well with human's perception of image quality. For this reason, many new metrics were proposed: among these there are SSIM [26] and MS-SSIM [27]. These metrics try to summarize the structural similarity between two images taking into account the property of the human vision system.

The vast majority of CNNs for image restoration use the ℓ_2 loss. The authors of [33] highlight the limitations of such loss in restoration tasks and propose alternatives based on the ℓ_1 error and the SSIM and MS-SSIM image quality indexes (which are differentiable). They find that a mixed loss, composed of a combination of MS-SSIM and ℓ_1, achieves the best performance in both traditional and perceptually motivated metrics. Moreover, they show that switching ℓ_2 to ℓ_1 (and vice versa) when the training loss stops decreasing, can significantly improve the final performance.

2.3. Discussion

In the last few years, the number of proposed deep learning models for image restoration, in particular image denoising, has dramatically increased. A large variety of architectures and approaches have been proposed and today there exist methods able to deal with different noises, both synthetic and realistic. However, it seems that few works have addressed in detail the role of the receptive field in CNN denoisers, specifically for what concerns the effective receptive field [17]. Another possible research direction is the study of modules or procedures to exploit rotation and translation equivariance inside the networks, possibly with the use of RNNs. A possible starting point is RotEqNet [18], which proposes CNN modules to deal with rotation invariance, equivariance and covariance inferences for high level vision tasks.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Buades, A., Coll, B., and Morel, J.-M. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 2, IEEE, pp. 60–65.

[3] Chen, C., Xiong, Z., Tian, X., and Wu, F. Deep boosting for image denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–18.

[4] Cruz, C., Foi, A., Katkovnik, V., and Egiazarian, K. Nonlocality-reinforced convolutional neural networks for image denoising. *IEEE Signal Processing Letters 25*, 8 (2018), 1216–1220.

[5] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing 16*, 8 (2007), 2080–2095.

[6] Foi, A., Trimeche, M., Katkovnik, V., and Egiazarian, K. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing 17*, 10 (2008), 1737–1754.

[7] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

[8] Guo, S., Yan, Z., Zhang, K., Zuo, W., and Zhang, L. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 1712–1722.

[9] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[10] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (2015), pp. 448–456.

[11] Katkovnik, V., Foi, A., Egiazarian, K., and Astola, J. From local kernel to nonlocal multiple-model image denoising. *International journal of computer vision 86*, 1 (2010), 1.

[12] Kokkinos, F., Marras, I., Maggioni, M., Slabaugh, G., and Zafeiriou, S. Pixel adaptive filtering units. *arXiv:1911.10581* (2019).

[13] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[14] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning* (2018), pp. 2965–2974.

[15] Lemarchand, F., Nogues, E., and Pelcat, M. Noisebreaker: Gradual image denoising guided by noise analysis. *arXiv:2002.07487* (2020).

[16] Liu, D., Wen, B., Fan, Y., Loy, C. C., and Huang, T. S. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems* (2018), pp. 1673–1682.

[17] Luo, W., Li, Y., Urtasun, R., and Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems* (2016), pp. 4898–4906.

[18] Marcos, D., Volpi, M., Komodakis, N., and Tuia, D. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 5048–5057.

[19] Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision* (July 2001), vol. 2, pp. 416–423.

[20] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch.

[21] Plotz, T., and Roth, S. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1586–1595.

[22] Remez, T., Litany, O., Giryes, R., and Bronstein, A. M. Class-aware fully convolutional gaussian and poisson denoising. *IEEE Transactions on Image Processing 27*, 11 (2018), 5707–5722.

[23] Tai, Y., Yang, J., Liu, X., and Xu, C. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 4539–4547.

[24] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. Deep learning on image denoising: An overview. *arXiv preprint arXiv:1912.13171* (2019).

[25] Vedaldi, A., and Lenc, K. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), pp. 689–692.

[26] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612.

[27] Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (2003), vol. 2, Ieee, pp. 1398–1402.

[28] Yu, F., and Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

[29] Yue, Z., Yong, H., Zhao, Q., Meng, D., and Zhang, L. Variational denoising network: Toward blind noise modeling and removal. In *Advances in Neural Information Processing Systems* (2019), pp. 1688–1699.

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

[31] Zhang, K., Zuo, W., Gu, S., and Zhang, L. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 3929–3938.

[32] Zhang, K., Zuo, W., and Zhang, L. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing 27*, 9 (2018), 4608–4622.

[33] Zhao, H., Gallo, O., Frosio, I., and Kautz, J. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging 3*, 1 (2016), 47–57.