Self-supervised image denoising with deep neural networks

Stone Fang (Student ID: 19049045)

Abstract

Image denoising, as a fundamental task in computer vision (CV), has been intensively studied for decades. Also, it is an important processing stage for other CV tasks such as object detection. This project will conduct a in-depth study of self-supervised deep learning image denoising approaches, that is, deep neural models without the presence of clean targets in the same domain for model training. CNN-based methods will be focused, along with some improvements including residual learning and batch normalisation. Moreover, a GAN-based noise level modelling and estimation method will be employed as an adaptive way to specify the noise map input in FFDNet. In terms of self-supervision, meta-learning approaches will be conducted and combined with CNN-based models, which makes the proposed method able to learn on a large variety of datasets. To validate the model, experiments will be taken on a large variety of dataset including both synthetic and real word noisy images. The planned time table, resource requirement, expected outcomes, and possible risks are also described and discussed. Altogether, this study is expected to devise an effective and practical image denoiser able to work flexibly on various scenarios especially those under real-world environments.

1 Introduction

Removing noise from degraded images to recover high quality ones, known as image denoising, is considered as a fundamental area in computer vision. It has been a classic research topic yet remains active nowadays (Gu & Timofte, 2019). It not only greatly affects user experience in practical applications, but is also very important for subsequent computer vision tasks, for instance, classification and recognition (Gu & Timofte, 2019).

A widely accepted yet simple image degradation model is $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$ where $\boldsymbol{x}$ refers to the uncorrupted image, $\boldsymbol{y}$ represents the degraded image and $\boldsymbol{n}$ is the additive noise (Gu & Timofte, 2019; Zhang, Zuo, Chen, Meng, & Zhang, 2017). Several kinds of noises has been widely studied, including additive white Gaussian noise (AWGN), Poison noise, and salt-and-pepper noise (Gu & Timofte, 2019).

The biggest challenge in image denoising is the loss of information during degradation, making this problem highly ill-posed (Gu & Timofte, 2019; Lee, Cho, Kim, & Kim, 2020). As a result, prior knowledge is required to compensate the lost information to recover high quality image (Gu & Timofte, 2019). This can be the prior modelling of either the images or noise (Chen, Chen, Chao, & Yang, 2018). Based on the information used in modelling, image denoising methods are generally divided into two categories (Gu & Timofte, 2019): a) *internal*, which only use the noisy images; b) *external*, which use both noisy and clean (ground truth) images. These two approaches can be combined or mixed to reach better performance.

A variety of models have been proposed for image prior representation, including some state-of-art ones such as BM3D or WNNM, which are based on non-local self-similarity features (Valsesia, Fracastoro, & Magli, 2019; Zhang et al., 2017). The most popular and classic one is BM3D, which serves as a benchmark in image denoising (Chen et al., 2018). However, there are some major disadvantage of these models. First, they mostly rely on human knowledge. Second, they only utilise the information of a single input image (Chen et al., 2018).

In recent years, deep neural networks (DNNs) have revolutionised traditional methods and became the state-of-art technology on most tasks of computer vision (Gu & Timofte, 2019). In terms of image denoising, a range of DNN models have been proposed, attracting increasing attentions attributed to its performance. Models based on Convolutional Neural Networks (CNNs) achieved significance, such as RED (Mao, Shen, & Yang, 2016) or DnCNN (Zhang et al., 2017). More recent technologies are also introduced to image denoising, such as Generative Adversarial Networks (GANs) (Chen et al., 2018), Graph Neural Networks (GNNs) (Valsesia et al., 2019), and meta-learning (Lee et al., 2020). Self-supervision is one of these trends. It is also learning in a supervised way, that is, with input and label, but the labels are autonomously generated in the absence of human effort. In this manner, deep learning image denoiser can be trained only on noisy images, or even a single noisy input (Krull, Buchholz, & Jug, 2019; Lee et al., 2020).

2 Related Work

There are several CNN-based image denoising models. RED-Net was proposed in (Mao et al., 2016) for denoising in different noise levels with a single model. It is a very deep architecture

(up to 30 layers in experiment) which consists of convolutional and deconvolutional layers, with skip connections linking every convolutional layer with its symmetric deconvolutional one. The skip connections helps passing gradients in back-propagation to alleviate gradient vanishing problem. This model outperforms existing state-of-art models in image denoising, and is claimed to be the first approach with good metrics working at different noise levels with only one model. Another work of DnCNN (Zhang et al., 2017) combines residual learning strategy and batch normalisation into feed-forward convolutional neural network to improve the final model metrics and also to accelerate the training process. An advantage of DnCNN is that it learns the residual (that is, the difference between the clean and noisy image) instead of clean image itself because the image patterns are greatly more complex than noises. This model outperforms state-of-art methods such as WNNM, TNRD and BM3D, and can be effectively extended to more generalised image denoising tasks, for instance, blind Gaussian denoising. However, such model is only optimised for a specific noise level, and lack of flexibility for spatially variant noise. To solve these problems, Zhang, Zuo, and Zhang (2018) proposed FFDNet by introducing a noise map $\boldsymbol{M}$ along with noisy image as the input of CNN. In addition, FFDNet works on down-sampled sub-images not only to increase computational speed but to expand the receptive field. The experiments demonstrated that FFDNet is able to work on a large range of noise levels without re-training the model, as well as spatially variant noise. Its effectiveness was also shown by real world images. On the other hand, the noise map has to be manually set as the input of the model instead of a learnable parameter. It is also worth noticing that though DnCNN and FFDNet both outperformed traditional models BM3D and WNNM, they reached less performance than the latter on the Barbara image in SET12 dataset on at all noise levels (Zhang et al., 2018). Moreover, some recently emerged deep learning technologies are also applied successfully. For instance, Valsesia et al. (2019) employed a graph neural network to CNN-based architecture. Owing to the local nature of convolutional operation, CNN-based model is unable to exploit non-local similarity patterns which had been proven to be significant by previous model-based methods. As a response, GraphCNN was proposed by incorporating the Edge Conditioned Convolution (ECC), a graph convolutional layer, to create non-local receptive field. This method improved the metrics on average, but did not beat existing methods on some categories in their experiment.

Some attempts has been made to loosen the requirement for training data. Chen et al. (2018) proposed GCBD model, arguing that in real world applications noise is not ease to obtain and is usually more complex than Gaussian noise, thus the models trained for knowing noises might

significantly decrease in performance. To solve this problem, the GCBD utilise a two-stage model. First, a GAN learns to model the distribution of noises. Second, the noise sampled from previous step are paired with real images, resulting in a proper training dataset for deep CNN based denoising models. Though this method does not require noisy and clean image pairs for training, it still needs clean images. Another method named Noise2Noise (Lehtinen et al., 2018) learns a denoising model with noisy data only, based on an basic observation that the loss function will not be affected by the change of distribution of targets as long as it remains the same expectation. This model can be trained without accessing to clean images yet achieves comparable performance, if not better, of models trained from dataset with clean targets. The capability of Noise2Noise was demonstrated by experiments with various noises and images, but they only covered synthetic and real word noisy images. Another drawback of this method is that it needs different noisy observations of the same image. Going one step further, Noise2Void (Krull et al., 2019) implements image denoising with only a single noisy input, and can be applied to many existing neural network architectures. In this approach, the author introduce the blind-spot network, trained by patches extracted from noisy image with the center pixel masked. Experiments were conducted not only on synthetic noises but also on real noisy images, however, the results on real data were evaluated by human vision due to the absence of ground truth. A latest research (Lee et al., 2020) proposed a two-phase denoiser which first utilised an arbitrary pre-trained denoiser g and augmented the available patches at self-supervision stage by adding random noise to the output of g. To gain benefit from supervised learning on large labelled datasets, a meta-learning approach was employed for fast adaptation to test inputs. The distribution of training dataset was not necessarily be identical to the test input, enabling this model to utilise large amount of available image datasets to learn general knowledge.

Many image datasets can be used for denoising model evaluation with synthetic noises, such as Set14 (Zeyde, Elad, & Protter, 2012), BSD300 (Martin, Fowlkes, Tal, & Malik, 2001) and its newly extended version BSD500 (Arbelaez, Maire, Fowlkes, & Malik, 2011). Furthermore, some real world data are available as benchmark. For example, Darmstadt Noise Dataset (DND) (Plotz & Roth, 2017) contains 50 pairs of images captured by consumer cameras at different ISO values, with the low-ISO ones as ground truth. Another real image dataset is Smartphone Image Denoising Dataset (SIDD) (Abdelhamed, Lin, & Brown, 2018) containing 30,000 noisy images.

3 Methodology

This research will be conducted based on several state-of-art approaches and experiments. In addition to the improvement on model architecture, this study will also focus on self-supervised training to minimise the dependency on in-domain training data. The overall architecture is shown in Fig 3. The datasets and evaluations used for this study will also be described.

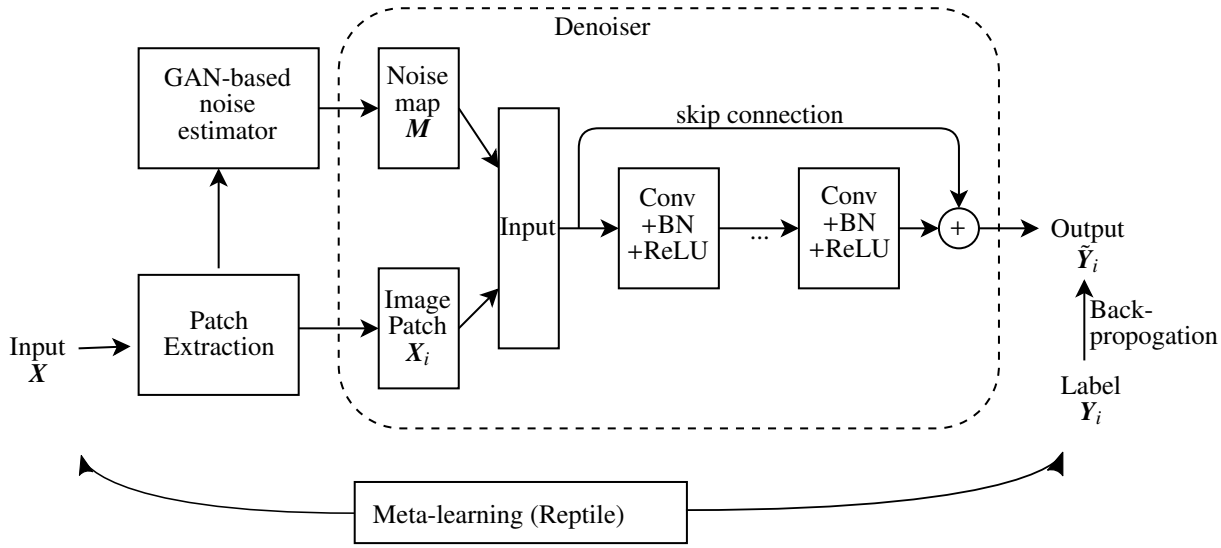

Figure 1: Overall architecture of image denoising model

3.1 Dataset

This study will do experiments on some commonly used dataset for image denoising task such as Set14 (Zeyde et al., 2012) and BSD500 (Arbelaez et al., 2011). Moreover, model performance in real world environments is an important goal of this study, so datasets of real noisy images will be incorporated, for instance, DND (Plotz & Roth, 2017) or SIDD (Abdelhamed et al., 2018). In addition, thanks to the meta-learning based approach which is able to transfer knowledge across domains, more image datasets can be employed as extensions for training data, such as ImageNet.

3.2 Neural Network Architecture

Several existing studies have given us useful guidelines in choosing model architecture. First of all, CNN has been proved to be effective and successful in image denoising (Mao et al.,

2016), so this study will focus on CNN-based approach. Second, residual learning and batch normalisation are great methods to improve the final model metrics (Zhang et al., 2017), so they will be employed by this project. Thirdly, FFDNet (Zhang et al., 2018) has demonstrated that the extra input of noise map can improve the flexibility and adaptation under different noise levels, so it is also taken into account. However, the noise map is trickily specified rather than learned in original FFDNet. To make the noise level estimation more adaptive, a GAN-based noise modelling method (Chen et al., 2018) will be introduced as a solution for this part. The GAN-based model will automatically learn the noise distribution in the image and generate the noise map $\boldsymbol{M}$ accordingly, removing the human effort of specifying the noise map.

3.3 Self-supervised Adaption

Self-supervised training is another important aspect of this research. This study aims to effective and practical image denoising method without dependency on large in-domain labelled training dataset which is collected at high cost. For this purpose, self-supervised approaches such as Noise2Noise (Lehtinen et al., 2018) and Noise2Void (Krull et al., 2019) are promising candidates. However, training solely in noisy input is unlikely to beat the models trained on larger labelled in-domain dataset. meta-learning has been proved to be effective to improve the model metrics and speed up model inference (Lee et al., 2020), and thus will be investigated in this study. GAN-based noise modelling (Chen et al., 2018) is another measure to alleviate the lack-of-data problem, which can be introduced to the two-phase denoiser (Lee et al., 2020) for patch generation with more realistic noise distributions.

3.4 Evaluation

There are several measurements for image denoising evaluation, among which peak signal to noise ratio (PSNR) is the mostly adopted one. Assuming that the estimation image is represented as E and ground-truth as G both in dimension of $M \times N$, PSNR is defined as (Gu & Timofte, 2019)

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right)$$

where R is the maximum signal range, and MSE is mean squared error defined by (Gu & Timofte, 2019)

$$MSE = \frac{\sum_{M,N}[E(m,n) - G(m,n)]^2}{M \times N}$$

There is another type of metrics called perceptual quality measures, of which some representatives arestructural similarity (SSIM) and feature structural similarity (FSIM). But in this project PSNR is chosen as the primary measurement because it is used in most image denoising literatures (Chen et al., 2018; Krull et al., 2019; Lee et al., 2020; Zhang et al., 2017; Zhang et al., 2018).

4 Timetable

Task	Deadline
Final decision on the topic, create research questions	1 week
Literature review	3 weeks
Research proposal draft	1 week
Prototyping	4 weeks
First round of testing and analysis	4 weeks
Model improvement	4 weeks
Second round of testing and analysis	4 weeks
Write and present final results	4 weeks

5 Research Resources

Deep learning models are computational intensive, and recent models to be studied such as GAN, GNN, meta-learning are even more so. Proper hardware sets includes Intel i7 series CPU, an Nvidia Titan X series or 2080Ti GPU, and 1T disk storage. Tensorflow or Pytorch will be chosen as the modelling framework in consideration of many existing works having been implemented in either framework.

6 Planned Outcomes and Risks

The expected outcomes of this research are practical image denoising systems that are effective on real world environments and applications. It will utilise state-of-art technologies for novel solutions at this aim, and will improve the metrics (PSNR) on particularly real world image datasets such as DND (Plotz & Roth, 2017).

The biggest risk of this research is that the expected improvement cannot be achieved for certain methods. To reduce such risk, more methods could be included and combined, and flexibility of up to 4 weeks will be introduced into schedule.

References

Abdelhamed, A., Lin, S., & Brown, M. S. (2018, June). A high-quality denoising dataset for smartphone cameras. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1692–1700). doi:10.1109/CVPR.2018.00182

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 33*(5), 898–916. doi:10.1109/TPAMI.2010.161

Chen, J., Chen, J., Chao, H., & Yang, M. (2018, June). Image blind denoising with generative adversarial network based noise modeling. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3155–3164). doi:10.1109/CVPR.2018.00333

Gu, S., & Timofte, R. (2019). A brief review of image denoising algorithms and beyond. In S. Escalera, S. Ayache, J. Wan, M. Madadi, U. Güçlü, & X. Baró (Eds.), *Inpainting and Denoising Challenges* (pp. 1–21). Cham: Springer International Publishing.

Krull, A., Buchholz, T.-O., & Jug, F. (2019, June). Noise2void - learning denoising from single noisy images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2124–2132). doi:10.1109/CVPR.2019.00223

Lee, S., Cho, D., Kim, J., & Kim, T. H. (2020). Self-supervised fast adaptation for denoising via meta-learning. *arXiv preprint arXiv:2001.02899*.

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

Mao, X.-J., Shen, C., & Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2810–2818). NIPS'16. Barcelona, Spain: Curran Associates Inc.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). 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* (Vol. 2, pp. 416–423).

Plotz, T., & Roth, S. (2017). Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1586–1595).

Valsesia, D., Fracastoro, G., & Magli, E. (2019). Image denoising with graph-convolutional neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 2399–2403).

Zeyde, R., Elad, M., & Protter, M. (2012). On single image scale-up using sparse-representations. In J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, & L. Schumaker (Eds.), *Curves and Surfaces* (pp. 711–730). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017, July). Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155. doi:10.1109/TIP.2017.2662206. arXiv: 1608.03981

Zhang, K., Zuo, W., & Zhang, L. (2018, September). FFDNet: Toward a fast and flexible solution for CNN based image denoising. *IEEE Transactions on Image Processing*, *27*(9), 4608–4622. doi:10.1109/TIP.2018.2839891. arXiv: 1710.04026

Responses to Reviewers Comments on

Project Proposal

COMP825 2020 Deep Learning

Reviewer 1

Question 1.1: The abstract is too short.

Answer 1.1: The abstract has been elaborated.

Question 1.2: Methodology is not clear, a dataset is needed for this project using deep learning, and comparisons and evaluations should be provided.

Answer 1.2: Methodology has been improved, clarified, and better organised. Add descriptions of datasets to use and metrics for evaluations.

Reviewer 2

Question 2.1: DnCNNs does not generalize well to real noisy images and it works only if the noise level in image ranges is in a pre-set range.

Answer 2.1: The methodology has been improved by more flexible model of FFDNet which performs well under variant noise. Moreover, the incorporation of GAN and meta-learning can also increase the generalisation capability.

Question 2.2: FFDNet is an improvement to DnCNN which is flexible to noise levels and spatially invariant noise.

Answer 2.2: FFDNet has been critically reviewed in the literature review section, and its idea has been assimilated into the proposed methodology.

Reviewer 3

Question 3.1: There is no in-depth discussion of the advantages and disadvantages of various models and justification of chosen model.

Answer 3.1: Add more in-depth discussion of pros and cons of each study in the literature review. The justification and rationale of proposed methodology has been elaborated.

Question 3.2: Make it clear the difference between self-supervised and supervised learning.

Answer 3.2: The difference self-supervised and supervised learning has been clarified.