February 2023
Master's Degree Thesis

Multi-inference strategy for self-supervised denoising problem

Graduate School of Chosun University
Department of Computer Engineering
Nazmus Saqib

Multi-inference strategy for self-supervised denoising problem

자기지도 디노이즈 문제를 위한 다중 추론 전략

February 24, 2023

Graduate School of Chosun University

Department of Computer Engineering

Nazmus Saqib

Multi-inference strategy for self-supervised denoising problem

Advisor: Prof. Jung, Ho Yub, Ph.D.

A thesis submitted in partial fulfillment of the requirements for a master's degree

October , 2022

Graduate School of Chosun University

Department of Computer Engineering

Nazmus Saqib

This is to certify that the master's thesis of

Nazmus Saqib

has been approved by examining committee for
the thesis requirement for the master's degree.

사키브 나즈무스의
석사 논문 승인

위원장 조선대학교 교수 강문수 (인)
위 원 조선대학교 교수 정호엽 (인)
위 원 조선대학교 교수 김판구 (인)

2022년 12월

조선대학교 대학원

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Multi-inference strategy for self-supervised denoising

Nazmus Saqib
Advisor: Prof. Jung, Ho Yub, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

Self-supervised image denoising is a challenging problem that aims at signal reconstruction on a sparse set of noise measurements without any supervision of clean ground truths. Conventional supervised methods consider the noise recovery process as an ill-posed optimization problem with the availability of ground truth which is challenging in numerous domains. Self-supervised techniques alleviate the ground truth-unavailability issue by incorporating several complicated objective functions for proper noise removal and reconstruction. However, the diverse noise distribution of images is crucial for noise recovery. Moreover, to form a complex loss function, the methods need to rely on additional hyperparameters. However, optimal hyperparameter estimation is complicated, and any mistuning of the parameters results in over-smoothing and inconsistent structure recovery that is responsible for performance degradation. This paper proposes a self-regularization technique without using any hyperparameter to alleviate the aforementioned issues. Our multiple predictions acquired from a multi-inference self-supervised strategy are exploited as the regularization

parameters and produce a compact loss function. Moreover, the proposed self-regularized method achieves satisfactory performance using multiple models and follows a simple training strategy without any complexity. Our experimental results represent that our compact loss function can achieve satisfactory performances in comparison to other existing methods for both synthetic and real noise domains. We also implement our algorithm on practical applications to represent how such low-level vision task is effective in high-level vision applications. We represent a comparison scenario with weakly and un-supervised denoising methods to highlight our improved performance in the above applications.

한 글 요 약

자기 지도 디노이즈 문제를 위한 다중 추론 전략

사킵 나즈머스
지도교수: 정호엽
컴퓨터공학과
조선대학교 대학원

자체 지도 이미지 노이즈 제거는 깨끗한 실제 값의 지도 없이 노이즈 이미지 세트를 사용하여 깨끗한 이미지 재구성을 목표로 하는 ill pose 문제이다. 실제 값에 과대한 의존은 과적합 및 분산 감소와 같은 문제를 초래할 수도 있어서, 현대 접근 방식은 특정 데이터 확대 또는 정규화 기술을 사용하여 이러한 문제들을 우회한다. 그러나 이런 접근 방식은 여전히 다양한 영역에서 일반화된 성능을 달성하지 못하고 있다. 게다가, 이런 전략들은 노이즈 제거와 같은 분야에서 추가적인 하이퍼 파라미터에 의존한다. 최적의 하이퍼 파라미터 추정은 어려우며 파라미터를 잘못 조정하면 성능 저하를 불러일으킬 수 있는 over-smoothing 및 일관성 없는 이미지 구조가 발생할 수 있다. 본 논문에서는 앞서 서술한 문제점을 완화하기 위한 파라미터에 의존하지 않는 self-regularization 기법을 제안한다. 다중 추론 전략에서는 다양한 이미지 예측 결과를 정규화로 활용하여 콤팩트한 손실 함수를 정의한다. 더불어, 제안된 자체 정규화 방법은 모든 데이터 확대 기술과 모델 의존성에서 의존하지 않고 간단한 훈련 전략을 사용한다. 다양한 실험을 통해 제안 하는 손실 함수는 합성 노이즈 영역 및 실제 노이즈 영역에서도 기존 기법들과 성능 우위 격차를 보여주고 있다.

I. INTRODUCTION

Artificial intelligence has recently achieved tremendous progress through learning from massive amounts of carefully labeled data. This traditional learning strategy is known as supervised learning. However, collecting enough labeled data for classification is challenging and time-consuming in several tasks like image segmentation, object detection/localization, and person identification. For example, if any model is trained to perform as a translator for low-resource languages. Laterally, in low-level image regression-based tasks like image denoising, low-light enhancement, super-resolution and image restoration, collecting clean ground truths is costly in diverse domains like medical imaging and scientific learning. So, the burning question is how far AI will go with supervised learning facing these difficulties.

To answer the question, the AI community thinks about establishing common sense in machines like humans and animals in a real-life scenarios, which pushes them to learn new skills without requiring massive amounts of knowledge for every single object or task. At this stage, self-supervised learning is one of the most promising ways to acquire such background knowledge and common sense in AI systems. The prime criteria of self-supervised learning are twofold: the initial criteria is forming generalized predictive models enriched with learning concepts about the objects of the world. Secondly, building a working hypothesis (objective function) with trial and error which can explain how much the model learns from the environment.

Already the vision community has achieved tremendous progress in high-level vision-based tasks. They have launched a billion-parameter open-source computer vision model SElf-supERvised (SEER) [1], trained on numerous

amounts of unlabeled random public Instagram images. Recently, self-supervised learning has also achieved satisfactory performance in low-level vision tasks considering the clean ground truth unavailability issue. This thesis will discuss about a self-supervised learning strategy in image denoising, one of the major sub-domains in the low-level vision field.

A. Conventional supervised learning in image denoising

Denoising is a signal-processing method that reconstructs the signal from corruption by preserving useful information. Signal reconstruction on controlled and uncontrolled noisy measurements is an essential yet challenging task. Typical instrumental complexities and rough environments corrupt images with non-stationary noise, which is crucial in numerous vision-based applications like semantic segmentation [2], super-resolution[3], and object detection[4]. Even the recent virtual world demands high-quality noise-free images captured by mobile phones and relevant devices.

With the tremendous progress of Deep Neural Networks (DNN), deep-learning methods [5]–[10] remove synthetic noise more frequently than the conventional denoising methods like [11], [12]. However, the very earlier models were incompatible with diverse unknown noise distributions. This phenomenon occurs due to the large domain discrepancy between the training and test image noise distributions, which results in unreliability and in-applicability in practical applications. Later, single-model blind denoisers [5], [13], [14] are proposed to alleviate the issues trained by manually composing training set images with diverse noise distributions. However, learning diverse noise distributions is more complicated than learning simpler ones. Moreover, blind prediction of noise cannot retrieve the noisy images' underlying structures and textures. Considering

the issue, a few flexible networks [8], [15] are proposed based on the noise level estimator. These networks can deal with diverse noise distributions by exploiting additional information. However, the networks require more than a single stage due to the noise level estimator and fall in performance degradation when the estimator needs to estimate accurate information. Recently, several model-based deep learning methods [16]–[18] propose a combination of multiple sub-networks where each sub-network plays an individual contribution. Thus, the methods full-fill multiple purposes like over-smoothing and artifact removal in the resultant restored image, which are the main prerequisites for image denoising. However, the contribution of multiple sub-networks designed for solving multiple non-linear sub-problems achieves handsome noise recovery. Still, they require multiple complex functions and equations, which increases the computational burden.

Though supervised networks with different innovative strategies can achieve satisfactory performance in the synthetic domain, denoising in real-world applications is still challenging. Images captured by digital cameras or relevant devices comprise "real noise," which distribution experiences a significant deviation from the i.i.d. Gaussian distribution. In detail, issues of dealing with real-noise can be described in twofold. The primary concern is collecting noisy/clean pairs in real-world scenarios is challenging due to several complicated features like signal dependency, multi-modal, spatially variant, etc. Such complicated features create a challenging noise-recovery scenario during accurate modeling of real-world noise [19]–[23]. Few methods follow a costly solution of collecting vast amounts of noisy/clean image pairs with careful image acquisition settings [2,33] to alleviate the data scarcity problem. The remaining methods use either a Generative Adversarial Network (GAN) for data generation

or any prior knowledge of camera parameters and noise properties [16,5,42]. Secondly, applying the supervised strategy, it is inconvenient to learn such complex distribution of real noise using a single convolutional neural network (CNN). Few researchers propose precise real-noise modeling, but they suffer from textural loss due to relying on blind denoisers.

B. Self-supervised learning in image denoising

As a result, researchers follow the self-supervised learning strategy in image denoising, where no clean ground truths are required. These methods are unable to achieve outstanding performance due to the unavailability of clean ground truths but resolves the data scarcity problem and the non-negligible performance gap between synthetic and real noise domains. The self-supervised strategy first introduces with Noise2Noise(N2N) [24], which is the first "no-need ground truth" approach that learns a corrupted image and produces an estimation of the corresponding clean counterpart. However, collecting two noisy independent realizations from the same scene is quite difficult in dynamic scenarios where the variation of image quality is a continuous incident. To alleviate the above issue, a series of self-supervised approaches release the requirement of individual noisy observations. They require only a single image to produce the input-target pair to train the network. These studies can be divided into two categories regarding their training strategy. The initial category follows the same Noise2Noise framework to train their model and follows the pair-generation strategy with simulated noises using the known noise model. The second category is mask-based blind spot denoising, where only individual noisy images are available without extra information.

Conventional masking-based schemes [25]–[27] considers the central pixel as the blind spot on a larger receptive field. The strategy directs a standard self-supervision but losses the important context of the input image while the regressor selects a more extensive region to predict the blind spots. The textural loss results in poor performance in unknown noisy measurements. The performance degradation directs the way to either analyze the masking property in depth or a transition from invisible blind spots to visible ones. Through analyzing the masking-based property in depth. Few approaches use manual blind spot convolutional networks with post-processing like Bayesian approximation [28]–[30]. However, these methods still need to improve on information loss. Recently, Blind2Unblind (B2UB) [31] introduced a transition strategy from blind spots to visible ones that can reduce information loss. The improvement of masked-based approaches is effective in denoising and context retrieval but still requires raw RGB images to perform denoising. Usage of Raw RGB image is entirely irrelevant as the images are device dependent, specific to the corresponding camera parameters. Moreover, the approaches could be more generalizable as the mapping from raw RGB images can not achieve satisfactory performance in sRGB image denoising.

In contrast to the masked-based approaches, classical unblind methods [32], [33] follow several data augmentations for pair generation using multiple noisy realizations from a single noisy image. Their construction is based upon a specific noisy distribution, i.e., AWGN, to generate feasible input-target pairs. Although the pair-generation strategy is quite simpler than the previous masking approaches, data augmentation in such ways is quite complicated in the process. Adding fixed synthetic for pair generation is not generalizable to real domains due to non-negligible domain gaps between synthetic and real. Usually, real

noise is the consequence of an inevitable photon fluctuation occurring on the camera sensor. The noise follows the Poisson distribution, where the variance is proportionally dependent on the mean intensity at a specific pixel and not stationary over the whole image. However, contemporary unblind approaches train their model using stationary noise generally assumed i.i.d Gaussian, which is contradictory in approach. Therefore, training the regressor against heteroscedastic Gaussian noise is identical since the variance of this distribution is intensity-dependent. In the same way, we revisit the traditional pair-generation strategy due to training simplicity and consider inherent noise as variable Gaussian noise instead of noise level estimation with known noise prior. Our pairing strategy conforms that the additional synthetic noise for the training purpose will be inherent noise independent, which is preferable in both synthetic and real noise domains. From this scenario, our intuition is to estimate close to the clean image by observing the double noisy observations of the single noisy image.

Moreover, contemporary approaches assume a family of losses [31], [34] instead of a single loss function while thinking of a model with multiple qualities. For example, while training any image restoration model, the practitioner often thinks about some good prediction corresponding to the target and joint optimization of the size of restored images and their visual quality. In such scenarios, a simultaneous minimization of multiple loss functions is required, where each of them will consider a different facet of the problem mentioned above.

Typically, designing a compact loss function is complicated due to the model capacity, where all inherent losses are in simultaneous optimization. Thus the designer has to decide about the balance of the intrinsic losses

during optimization. Existing approaches scalarize such multiple objective functions by linearly combining the inherent losses with single or multiple additional hyperparameters that can define the trade-off between the loss terms. However, the values of the additional hyperparameters strongly affect the model's performance, and tuning the hyperparameters can be cumbersome. Moreover, it might cause uncertainty about how the hyperparameter tuning affects the final values of the overall compact loss functions. Furthermore, such a hypothesis cannot achieve generalizable performance in synthetic and real domains. Therefore, we propose a novel objective function with a regularization term independent of any additional hyperparameters.

Overall, we design a novel multi-stage sequential inference strategy established self-supervised. We assume that our consecutive predictions from the sequential stages contain diverse noise observations, and by learning the diversity, the model will be able to produce an optimal solution of clean image estimation by observing extremely noisy observations. We propose a compact loss function combined with an additional regularization term to utilize the properties of multiple predictions. The multiple regularization terms of the compact loss function restrict the method to produce a satisfactory estimation by avoiding over-smoothness with inconsistent textural recovery. The overall motivations and contributions of our proposed method can be written as follows:

C. Motivations

Despite the tremendous progress of alternate supervision application in denoising domain, the performance of unsupervised learning methods for denoising is still not compatiable to the existing supervised methods like Feed-forward Denoising

Convolutional Neural Networks (DnCNN) [5] trained on noisy/clean pairs or the weakly-supervised method Noise2Noise (N2N) [24] trained on noisy/noisy pairs. Indeed, their performance is different from the classical nonlocal denoising methods such as Block-matching and 3D filtering (BM3D) [11]. Considering the data-unavailability issue, the weakly supervised method is not ethical if we require no clean ground truth in this domain. In summary,

- Unsupervised learning performs satisfactorily in real-world applications since it remains useful when no-ground truth image is available.
- Most existing unsupervised learning methods have a noticeable performance gap to their supervised counterparts, especially for denoising real-world images.

We are greatly inspired by the tremendous success of unsupervised methods where the unavailability of ground truth is the major issue. However, these methods' performance is not generalizable for some reasons. Like these, a few fundamental issues of the noise recovery process motivate us to find a new way. We can briefly describe the issues as follows:

- **Inherent noise dependency:** We follow the pair-generation strategy by adding synthetic noise inspired by a few unblind methods [32], [33]. According to their methods, the extra addition of synthetic noise depends on the unknown amount of inherent noise. As a result, this strategy is questionable.
- **Hyper-parameter dependency in regularization term:** Contemporary approaches [31], [34] propose objective function with additional

regularization term. However, the regularization term seeks additional hyperparameter tuning to stabilize the training procedure. Such a rigorous strategy suffers from non-generalizable performance in diverse domains.

- **Performance gap between synthetic and real domains:** The basic requirement of a self-supervised denoising method is to achieve a generalizable performance in both synthetic and real-domains. Most of the existing methods are established on either synthetic or real domain.

The following issues motivate us to find an inherent-noise independent pair-generation strategy and design a loss function estimation without using any hyperparameter. The proposed method not only provides the SOTA performance among existing unsupervised learning methods, but also is very competitive to many supervised learning methods including DnCNN.

D. Contributions

Considering all of the prerequisites of existing methods, we propose a framework which contributions can be described as follows:

- We propose a simple yet efficient pair-generation strategy that can mitigate the inherent noise issue and bounds to remove the non-negligible gap between synthetic and real noise domains.
- Our multi-inference strategy produces multi-predictions with diverse noisy observations force the model to solve a counterintutive approach where observing two double noisy images can find a way of clean image approximation.

- Our compact loss function is established with an additional regularization term which is completely hyperparameter-dependent that restricts the method from fulfilling the basic requirements of denoising, like avoiding over-smoothness and inconsistent textural recovery.
- Experimental results demonstrate that our method achieves generalizable performance among several un-/self-supervised methods on both synthetic and real noise domains.

E. Thesis Layout

The thesis follows a sequence: Section II describes the theoretical background of self-supervised denoising. Section III describes the studies related to our proposed method. The main section is section IV which describes our method. The next section provides different experiments to prove our method is effective. The last section concludes with a conclusion.

II. Background

This section will provide a theoretical overview of supervised to self-supervised methods. Moreover, a brief explanation will be discussed about the regularization hypothesis which is implemented by the existing self-supervised denoising methods.

A. From supervised to self-supervised

Let us consider a supervised training scenario where x is the available ground truth for any noisy measurement y. Training a regressor model $f(\cdot)$, parameterized by θ, implements the following empirical risk minimization:

$$\ell_{sup} = \mathbb{E}_{x,y}||f_\theta(y) - x||_2^2 \tag{1}$$

Such a strategy is adequate to achieve outstanding denoising performance, but the acquisition of paired clean images is impossible in real-world applications.

Therefore, contemporary self-supervised methods eliminate x with $\hat{y}$ and apply a variant of loss functions whose general representation is as follows:

$$\ell_{self} = \mathbb{E}_{y}||f_\theta(y) - \hat{y}||_2^2 \tag{2}$$

where the target $\hat{y}$ is a modification of the input y, depends on the method's individuality. Noise2Noise (N2N) [24] replaces $\hat{y}$ with another noisy realization of the same scene of the input image y, which turns the method into a weakly-supervised training strategy. However, collecting short exposure pair $(y, \hat{y})$ from the same scene of an image is challenging in several domains like Flow Cytometry devices. As a result, researchers use a single noisy image y for their training procedure.

1. Blind-spot based methods

Masked-based methods [25]–[27] replace y with customly designed masker volume as an implementation of blindness where the masker volume is considered as either value from the neighborhood or any random ones.

The initial representation of mask-based methods is Noise2Void(N2V) [25]. This method assumes a receptive field y_{RF} in the single noisy image y, keeping the blind spot at its center, and the target $\hat{y}$ is replaced with a randomly selected value from the surrounding area of the center pixel. They trained their denoiser by minimizing the empirical risk as follows:

$$\ell_{n2v} = \mathbb{E}_y ||f_\theta(y_{RF}) - \hat{y}||_2^2$$

The second representation is Noise2Self(N2S) [26], which introduces $\mathscr{T}$-invariant mask to ignore the center pixel. The denoiser predicts the value into the mask y_J using the values outside of the mask J^c. Employing a $\mathscr{T}$-invariant function $f(y_{J^c})$, the method proposed a self-supervised loss which can be expressed as follows:

$$\ell_{n2s} = \mathbb{E}_J \mathbb{E}_y ||f_\theta(y_{J^c})_J - y_J||_2^2$$

However, Noise2Same(N2Same) [27] analyzes that the $f(\cdot)$ still shows weak dependency on values on J and thus does not strictly satisfy the $\mathscr{T}$ invariance property. To mitigate the mutual influences among the locations within J, they sample the values within J and proposed a self-supervised loss function $f(\cdot)$ to be strict $\mathscr{T}$ invariant. If J contains m number of values, then the minimization loss function can be expressed as follows:

$$\ell_{n2same} = \mathbb{E}_y||f(y) - y||_2^2/m + \gamma_{inv}\mathbb{E}_J[\mathbb{E}_y||f(y)_J - f(y_{J^c})_J||^2/J]^{1/2}$$

The initial term is the reconstruction loss and the second term is the invariance Mean Square Error (MSE). The invariance loss implicitly controls how f should be strict and should be $\mathscr{T}$ invariant. This is the basic difference between this method to the previous ones. The invariance term can control the f for being $\mathscr{T}$ invariant without any assumption about the noise requirement on f. This property allows the method to show a satisfactory performance with unknown noise models, inconsistent noise, or combined noise with different types.

However, the operation of the masker volume results in a sizable loss of valuable information. Continuous pixel value replacement with the masker value gradually decreases the original information from the noisy input image.

Considering this issue, Blind2Unblind(B2UB) [31] applied global-aware mask mapper $g(f_\theta(\Omega_y))$ on the masker volume of y, Ω_y for global denoising to reduce the information loss. The global mapper samples denoised volumes at blind spots and projects them onto the same plane to generate denoised images. The training strategy is using following loss function:

$$\ell_{b2ub} = \mathbb{E}_y||g(f_\theta(\Omega_y)) - y||_2^2 \tag{3}$$

However, the method is not generalizable due to requiring raw-RGB images. The method requires raw RGB images for training which can not show satisfactory performance for real-world images in sRGB space. Therefore, a standard approach is required to achieve a generalizable performance in all image spaces.

2. Unblind methods

In contrast, few unblind methods propose data augmentation by adding synthetic specific noisy distribution i.e. to generate input-target pairs. Noisier2Noise(Nr2N) [32] and NoisyasClean(NC) [35] consider a noisier image as input. The modification is synthesized by adding the noise β with the input image y.

$$\ell_{Nr2N} = ||f_\theta(y+\beta) - y||_2^2$$

Concretely speaking, y contains an unknown amount of inherent noise n from a known noise distribution. Noisier2Noise draws an additional synthetic noise sample β from the same noise distribution to add to the original noisy image y. Their intuition is to estimate the corresponding clean image of y through observing $y+\beta$ and y.

On the other hand, Recorrupted2Recorrupted(R2R) [33] proposed a rigorous pair generation strategy with a known noise level that is statistically equivalent to the Noise2Noise(N2N)[24]. This loss function can be expressed as follows:

$$\ell_{R2R} = ||f_\theta(y+\beta) - (y-\beta)||_2^2$$

Neighbor2Neighbor(NBR2NBR) [34] introduces a novel pair-generation strategy using subsamplers of the input image. Since η_1 and η_2 are the two neighbor subsamplers from y, which are considered as the input-target pair, the risk minimization equation can be expressed as follows:

$$\ell_{NBR2NBR} = ||f_\theta(\eta_1) - (\eta_2)||_2^2 \tag{4}$$

All of these modifications of pair constrict the methods to prevent the network from converging to a trivial identity mapping.

B. Self-regularization effect on loss function estimation

Certain self-supervised methods apply particular regularization techniques to penalize their model flexibility. Conventional blind-spot techniques [25], [26] overcome over-fitting issues through regularization. Self2Self(S2S) [36] introduces dropout that can provide an efficient training strategy by reducing large variance caused by a single training sample. Recent approaches proposed additional regularization terms with empirical risk minimization that is controlled by a hyperparameter. The combined loss function intends to relieve the denoiser from converging to an identity mapping. Specifically, the general loss function with additional regularization terms can be expressed as follows:

$$\ell = \mathbb{E}_y ||f_\theta(y) - \hat{y}||_2^2 + \gamma.\mathbb{G} \tag{5}$$

Where $\mathbb{G}$ is the additional regularization term controlled by a hyper-parameter γ. Contemporary approaches [31], [34] replace $\mathbb{G}$ with another minimization term. Blind2Unblind(B2UB) [31] substitutes $\mathbb{G}$ with a risk minimization to constrain the blind term and perform the training without blind spots. The regularizer assists the method in retrieving useful information reduced by the conventional blind-spot methods. Similarly, Neighbor2Neighbor(NBR2NBR) [34] replaces the $\mathbb{G}$ with a minimization term to increase the training strength. The hyperparameter γ in this method performs as a controller between noisiness and smoothness. However, usage of such regularization not only depends on the hyperparameter extensively but also introduces additional hyperparameters that increases computational complexity.

Another concept of training scheme Stein's unbiased risk estimate(SURE) [37] proposed a regularization term that corresponds to the divergence with respect to y. Later, due to the complexity of divergence calculation, Monte Carlo Simulation is used [38]. However, additional hyperparameters result in more complexity. Moreover, the loss function deals with Poisson noise only. For Gaussian noise, when the noise variance varies, the model has to be trained again, which causes an additional computational burden.

III. Related Studies

Theoretical explanation and experimental results consist of several related works, specially the existing self-supervised denoising techniques that are applied in different domains. Our approach has inspected the intersection of multiple methods prevailing in different domains.

A. Non-learning based image denoisers

The earlier non-learning-based methods exploit manually design filters [39], [40] and perform an iterative filtering scheme to remove the image noise. Observing the counterpart scheme of a natural image patch, several methods perform the blockwise operation based on the spatial information [11], or non-local self-similarity [41]. Later, the remaining non-learning methods consider the denoising task as a formulation of a maximum posteriori (MAP) based optimization problem, whose performance depends on image priors. They exploit image patch as a sparse representation of the proper mathematical function and impose certain image priors like sparsity prior [42], [43], low-rank approximation [44]–[47]. However, the non-learning-based methods with strong mathematical derivations significantly suffer from structure loss and increase the computational complexity with iterative optimization.

B. Supervised learning with paired noisy/clean version

The supervised denoising methods learn latent mapping from the noisy/clean pairs. After the prominent success of Deep Neural Networks (DNN) in comparison to other manual algorithms, denoising methods proposed complex network architectures using feature attention [17], N3-block [10] for well-

generalized denoising performance. On the other hand, some deep unfolding-based methods [16] implement some traditional strategies like non-linear reaction-diffusion [6] and the proximal gradient method [48] by deep networks. However, the supervised methods exhibit several issues. The primary concern is model dependency. With an effective internal design of the network architecture, it is easier to show a well-generalized performance. Secondly, in hostile environments, the availability of clean counterparts is quite challenging.

C. Denoisers trained with pairs

Considering the unavailability issue of noisy/clean pairs, several deep denoisers use secondary noisy observations in replacement of the clean target. Noise2Noise(N2N)[24] achieves satisfactory performance using the double noisy pairs of the same scene. The such performance demonstrates that training any network following this manner can be able to assume an equivalent approximation of the clean target.

D. Unsupervised denoisers

Although pair generation with multiple noisy observations is more feasible than the noisy/clean image pairs, in several existing scenarios like CC/surveillance camera and medical imaging; obtaining multiple views from the exact same scene is way much more difficult. A few unsupervised methods address the issue. For example, Che *et al.* [49] generated synthesized noisy images to alleviate the problem. Few approaches [33] used unorganized noisy images instead of the organized noise pairs and designed a self-prediction loss function for Neural Network (NN) training. However, the convergence issue has become the main

concern for such approaches. Few unsupervised approaches use Generative Adversarial Network (GAN) for training pair generation [49], [50] or directly train the deep denoiser on noisy images. However, the unsatisfactory performance of the above methods and the real-world data availability led the researchers to find a self-supervised way where the feedback provided by each sample is huge.

E. Self-supervised denoisers

Noise2Noise (N2N)[24] is the first representational approach that requires two independent noisy observations without any clean counterparts. Later, mask-based approaches [25]–[27] are introduced where masking on a single noisy image, the denoiser can produce an approximation of the corresponding clean image. However, the masking strategy causes an unexpected information loss due to replacing the pixels with blind spots. Therefore, few approaches transfer the masking strategy to either a larger receptive field [28], [29] or global denoising [31] for reducing the information loss. However, the computational complexity of the above methods introduces the pair generation strategy instead of masking. These types of methods [32], [34] propose different data-augmentation to avoid over-fitting and use the traditional Noise2Noise framework between them. For better accuracy and stable training purpose, a few of the above methods introduce different additional regularization term [31], [34], [36]. Our method falls into a similar category. Table 1 represents a brief review of the existing weakly supervised, unsupervised and self-supervised methods in image denoising.

Category	**Method**	**Training image pair generation**	**Risk minimization and regularization**
Noisy/clean	Noise2Noise(N2N)[24]	noisy/clean	Conventional ℓ_2 norm minimization
Multiple noisy realizations from a single image	Noisier2Noise(Nr2N)[32]	Generates a synthetic noise sample, add it to the already noisy image which is considered as the input, the target is the noisy image	ℓ_2 norm minimization between input and target and estimate the clean image.
	Recorrupted2Recorrupted(R2R) [33]	Corrupts both input-target pairs with known noise levels.	ℓ_2 norm minimization.
	Neighbor2Neighbor(NBR2NBR) [34]	Generates a pair of sub-sample images from a single noisy image and consider them as input-target pair	ℓ_2 norm minimization between the subsamplers.
Masked based blind-spot methods	Noise2Void [25]	Implements masking on the input image and the masked values are replaced with select random values from the local neighbors.	ℓ_2 norm minimization .
	Noise2Self(N2S) [26]	Masked values are replaced with some random values.	ℓ_2 norm minimization .
	Self2self (S2S)[36]	Pair generated by the Bernoulli sampler .	ℓ_2 norm minimization and drop-out regularizer .
	Noise2Same (N2S) [27]	Propose a new self-supervised loss without any extra information.	ℓ_2 norm minimization.
	Blind2Unblind (B2UB)[31]	Performs global denoising by global-aware mask mapper and transits from blind to unblind.	ℓ_2 norm minimization with re-visible loss as a regularizer.

Table 1: A brief review of the existing self-supervised methods which are categorized as their proposed strategy.

IV. Proposed Framework

The initial subsection of the methodology represents an intuitive description and further mathematical justification of our proposed method. The remaining describes the overall framework in a subsequent manner.

A. Intuition

Let us consider $\mathscr{A}$ as a known noise distribution. We have drawn three random samples n, M_1, and M_2 from $\mathscr{A}$. We observe the two sums $\mathbb{S}_1 = n + M_1$ and $\mathbb{S}_2 = n + M_2$. Our intuition is to consider $\mathbb{S}_1$ as the input and $\mathbb{S}_2$ as the target noise distribution and observing the two sums ask the model to predict n. There is no direct way to predict n by distinguishing the contributions of M_1 and M_2. One plausible solution is incorporating the squared error function to minimize the distance between model prediction and the target. However, predicting a less noisy distribution by observing two higher noisy distributions is a counterintuitive approach. The most common scenario is to predict any noisy distribution through producing a way from higher to lesser noisy distribution [32] or observing modified realizations of similar noise distributions [34]. If our approach over the training procedure can reduce an approximate amount of M_2 from the prediction of the model, it might be possible to reach a clean estimation. In this scenario, our prediction should be $E[n|\mathbb{S}_1]$, which can be also considered as symmetrically $\mathbb{S}_1/2$ or $\mathbb{S}_2/2$. As n, M_1, and M_2 are independently drawn from the $\mathscr{A}$, any estimate of n can be equal to the corresponding implicit estimate of M_1 or M_2. Considering these, if we assume two double noisy distributions at both input and target ends, we have to reduce a similar amount of $\mathbb{S}_2$ as well as twice the estimation of M_2 from the prediction of the model to reach the ultimate goal.

This is the insight of our method inspired from [32]. Our intuition can be

directly applied if we consider a complicated scenario of any unknown noise quantity x, and we can apply $x+\mathbb{S}_1$ as input and $x+\mathbb{S}_2$ as the target, we can still estimate $x+n$. Thus we can agree that it is possible to estimate a clean image through considering double noisier images as input and target.

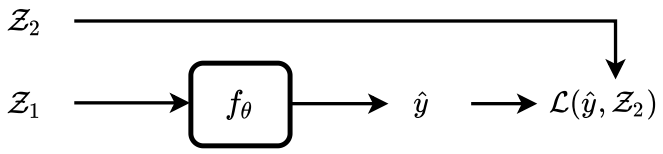

Figure 1: Illustration of our intuition. The training procedure for double noisy pairs $\mathscr{Z}_1, \mathscr{Z}_2$, the model gives the prediction $\hat{y}$, and through the loss function $\mathscr{L}(,\cdot,)$ minimizes the loss independently between $\hat{y}$ and $\mathscr{Z}_2$. Through the loss function, the implicit noise of $\mathscr{Z}_2$ instructs $\hat{y}$ about the amount of noise which is possible to reduce.

B. Mathematical justification

We consider the following learning scenario: given a distribution of natural images $\mathscr{N}$, and let $x \sim \mathscr{N}$. x is completely an unobserved distribution. However, if we consider a single noisy observation $\mathscr{Y} \triangleq x+n$. Here, $n \sim \mathscr{A}$, variable AWGN with the range $[5,25]$ where $\mathscr{A}$ is a known noise distribution. As $\mathscr{A}$ is known, we can draw two additional synthetic samples M_1 and M_2.

Our training procedure is as follows: from the given noisy image $\mathscr{Y}$, we produce two noisier versions, $\mathscr{Z}_1 \triangleq \mathscr{Y}+M_1 \triangleq x+n+M_1$, and $\mathscr{Z}_2 \triangleq \mathscr{Y}+M_2 \triangleq x+n+M_2$. We can consider a self-supervised training scenario where a given distribution of pairs $\mathscr{Z}_1, \mathscr{Z}_2 \sim P_{\mathscr{A}}$ with $\mathscr{Z}_1, \mathscr{Z}_2 \in \mathscr{A} \subset \mathbb{R}^{d_{\mathscr{A}}}$, and a loss function $\mathscr{L}(\cdot,\cdot) : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$, we aim to learn a model $f : \mathscr{A} \to \mathscr{A}$, with parameters θ, such that its predictions: $\hat{y} = f(\mathscr{Z}_1, \theta)$ minimize the expected value of the loss

$\mathscr{L}(\hat{y}, \mathscr{Z}_2)$ over the dataset.

It is impossible for the network to observe n or M individually and simply subtract them from the input $\mathscr{Z}_1$ to get the clean image. The possible strategy is to operate the noise distribution of any realization by observing another one. Thus, for given training distribution of pairs $(\mathscr{Z}_1, \mathscr{Z}_2)$; we can predict $E[\mathscr{Z}_2|\mathscr{Z}_1]$. Therefore, based on the intuition, it is possible to extract a clean image estimation $E[x|\mathscr{Z}_1]$, from an estimate of $E[\mathscr{Z}_2|\mathscr{Z}_1]$. If we recall that n, M_2 are two i.i.d noise samples alongside with M_1 and therefore $E[n|\mathscr{Z}_1] = E[M_2|\mathscr{Z}_1]$; using the relation we can explain our hypothesis as follows:

$$\begin{aligned} E[\mathscr{Z}_2|\mathscr{Z}_1] &= E[Y + M_2|\mathscr{Z}_1] \\ &= E[x + n + M_2|\mathscr{Z}_1] \\ &= E[x|\mathscr{Z}_1] + E[n|\mathscr{Z}_1] + E[M_2|\mathscr{Z}_1] \\ &= E[x|\mathscr{Z}_1] + E[M_2|\mathscr{Z}_1] + E[M_2|\mathscr{Z}_1] \\ &= E[x|\mathscr{Z}_1] + 2E[M_2|\mathscr{Z}_1] \end{aligned} \tag{6}$$

However, according to the equation 6, $E[\mathscr{Z}_2|\mathscr{Z}_1]$ will not be the exact reconstruction of the true clean image due to the addition of a noise term $2E[M_2|\mathscr{Z}_1]$ with the possible clean image estimation term $E[x|\mathscr{Z}_1]$. For the exact reconstruction of a clean image, we have to subtract the possible noise term $2E[M_2|\mathscr{Z}_1]$ from $E[\mathscr{Z}_2|\mathscr{Z}_1]$. We can express this mathematically as follows:

$$E[x|\mathscr{Z}_1] \triangleq E[\mathscr{Z}_2|\mathscr{Z}_1] - 2E[M_2|\mathscr{Z}_1] \tag{7}$$

Equation 7 justifies our intuition. Although M_2 is a known noise quantity, we are creating a training scenario where the model considers M_2 as an unknown quantity and reducing the term twice from the prediction, which is extracted

from $E[\mathscr{Z}_2|\mathscr{Z}_1]$. If the term $E[M_2|\mathscr{Z}_1] = 0$, then $E[\mathscr{Z}_2|\mathscr{Z}_1]$ can be an exact reconstruction of clean image estimation. $E[M_2|\mathscr{Z}_1] = 0$ is possible because M_2 is a sub-sample drawn from the zero-mean AWGN $\mathscr{A}$. During loss calculation, the value of M_2 is random every time, but the mean will always be zero. Thus the expectation of M_2 observing $\mathscr{Z}_1$ turns to zero and we can estimate the clean image while two double noisy images are considered as pairs for the training procedure.

Uncertainty of inherent noise distribution: According to [32], if we consider the Uncertainty of inherent noise distribution, we can assume the inherent noise is not from the same noise distribution of M_1 and M_2. In this scenario, we assume that $M_1, M_2 \sim \mathscr{A}$ and the inherent noise $n \sim \mathscr{B}$, where $\sigma_B < \sigma_A$. If both of $\mathscr{A}$ and $\mathscr{B}$ are zero-mean AWGN with $\sigma_B = \gamma\sigma_A$, then according to [32], $E[M_2|\mathscr{Z}_1] = \gamma^2 E[n|\mathscr{Z}_1]$. From that perspective we can rewrite 7 as follows:

$$E[x|\mathscr{Z}_1] \triangleq E[\mathscr{Z}_2|\mathscr{Z}_1] - 2\gamma^{-2}E[M_2|\mathscr{Z}_1] \tag{8}$$

The optimal value of γ can be either smaller or larger and depends on performance sensitivity. While the value of γ is 1, the inherent noise n can be from the same noise distribution of the additional Gaussian noise M_1 and M_2. The value of γ controls how much noise must be reduced during the training procedure. However, our equation does not depend on the value of γ while the term $E[M_2|\mathscr{Z}_1]$ turns to zero according to the above description. From that perspective, we can get a solution of such Uncertainty.

However, implementing such a counterintuitive approach with a single training stage like Fig.1 is cumbersome while the target is sufficient noise reduction and tends to find out a clean image through observing two higher

noisy distributions. It might be possible with a single training stage while the training pair exhibits very low noise distribution, and the target is to reduce a small amount of noise from the prediction. Considering both scenarios, the plausible solution can be employing multiple predictions of the model where each can act as an individual noisy realization. Thus, the training strategy can experience diverse noise distribution of the multiple predictions and finally produce an ultimate denoised image that will be close to the clean image as far as possible. Therefore, we propose a multi-inference strategy where the employment of multiple sequential predictions can observe diverse noise distribution during the training procedure. Simultaneously, we introduce an objective function with a regularization term. The regularization term consists of multiple individual objectives for an effective contribution of the multiple predictions in training.

C. Multi-inference strategy

This section provides a detailed description of our training strategy that uses the same denoiser f_θ multiple times to produce multiple predictions. Generally, self-supervised methods are model-adaptive and noise reduction performance is sensitive to the denoiser quality. As a result, existing methods using U-Net [2] architecture achieve better performance due to its multiscale operation and better reconstruction capability in comparison to the usage of traditional DnCNN [5] architecture. For our training, we use both DnCNN and UNet architecture to monitor our training strategy's independent performance.

The multi-inference strategy introduces four consecutive stages. The initial stage feds input noisy image $\mathscr{Z}_1$ to the denoiser and produces the primary

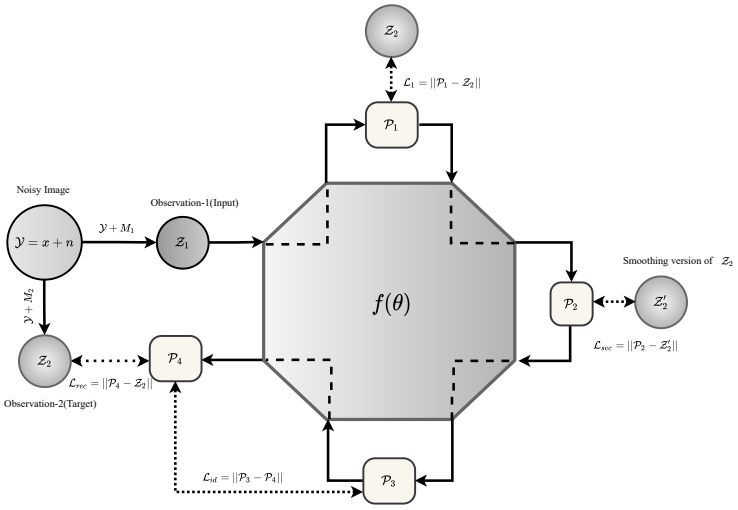

Figure 2: Semantic of our proposed framework. $\mathscr{Y}$ is the natural single noisy image and addition of M_1 and M_2 produces two observations $\mathscr{Z}_1$ and $\mathscr{Z}_2$. The octagonal represents the denoiser where $\mathscr{Z}_1$ is considered as input and $\mathscr{Z}_2$ is considered as target. The surrounding four points of the octagonal illustrates four predictions with corresponding loss functions.

prediction P_1 which can be expressed as $P_1 \triangleq f(\mathscr{Z}_1, \theta)$. Similarly, the next three consecutive stages produce three different predictions P_2, P_3 and P_4, which can be expressed as $P_2 \triangleq f(P_1, \theta)$, $P_3 \triangleq f(P_2, \theta)$, and $P_4 \triangleq f(P_3, \theta)$. Thus, the overall training procedure produces multiple noisy observations through which it can experience diverse noise distribution.

Finally, we can argue that to reach the target $\mathscr{Z}_2$ from the input image $\mathscr{Z}_1$, we experience four consecutive predictions. As a result, we can mathematically express the following observation:

$$E[Z_2|\mathscr{Z}_1] \triangleq E[P_1|\mathscr{Z}_1] + E[P_2|P_1] + E[P_3|P_2] + E[P_4|P_3] \tag{9}$$

The four individual terms define four sequential stages of inference. As the strategy is a consecutive decomposition procedure implemented by the same denoiser, the residuals of higher stages will suffer from over-smoothness or blurriness. The initial stages can experience better residuals with sharp edge preservation and fine details, but the sequential inferences will occur with an unexpected reduction of the high-frequency components. As a result, the very low-frequency signals of the image will cause random blurriness to the predictions at intermediate stages, which will hinder the prerequisite of a proper denoising method. Therefore, to experience an appropriate reconstruction, we incorporate several loss functions between the observations, which can participate from different perspectives.

D. Loss function

To train our network f_θ, we define a set of loss functions based on the statistical behaviors of the general noise. For every objective, we incorporate the L_2 loss that

can induce the mean-finding behavior between the prediction and target. Thus the loss can maintain the Uncertainty occurred by any unexpected imbalance between two noise distributions of the realizations.

Our initial approach is to introduce the traditional self-supervised loss like any self-supervised method for finding the halfway estimation between the prediction P_1 and the target image $\mathscr{Z}_2$. As P_1 and $\mathscr{Z}_2$ are two independent noisy observations, we tend to find the mean estimation between these two measurements through the self-supervised loss function. Our initial loss can be expressed as follows:

$$\mathscr{L}_1 = ||f_\theta(\mathscr{Z}_1) - \mathscr{Z}_2||. \tag{10}$$

where $||\cdot||$ is represented as L_2 norm for simplicity.

In the second stage, prediction P_1 is fed into the same denoiser and produces the prediction P_2. P_2 is a smoother version of P_1. Our secondary loss function is expected to predict a smoother estimation of the noisy version. Therefore, the loss function minimizes the distance between P_2 and the Fourier smoothing version of the target image $\mathscr{Z}_2$. Since the target image $\mathscr{Z}_2$ is considered in Fourier space, noise prevailing in the image contributes heavily to the high-frequency components. As a result, we reduce the amount of noise by applying the Fourier transform in $\mathscr{Z}_2$ image so that we can reduce the high-frequency components from the image, and thus we get the smoother version $\mathscr{Z}'_2$. Our secondary loss function can be expressed as follows:

$$\mathscr{L}_{sec} = ||f_\theta(P_1) - \mathscr{Z}'_2||; P_2 \triangleq f(P_1, \theta)) \tag{11}$$

In the third stage, two consecutive stages produce two intermediate assumptions P_3 and P_4. We construct the identity loss $\mathscr{L}_{id}$ based on the two above

inter-dependency assumption as follows:

$$\mathscr{L}_{id} = ||f_\theta(P_2) - P_4||; P_3 \triangleq f(P_2, \theta); P_4 \triangleq f(P_3, \theta) \tag{12}$$

However, the loss function has a chance to occur over smoothing to the final prediction. Therefore, we propose the reconstruction loss at the final stage. That minimizes the distance between the target and the final prediction P_4. We assume the general scenario where P_4 is supposed to be the smoothest and denoised version. Minimizing the distance between the target and the final prediction can retrieve the original information of the target image. The reconstruction loss can be expressed as follows:

$$\mathscr{L}_{rec} = ||P_4 - \mathscr{Z}_2|| \tag{13}$$

A combination of $\mathscr{L}_{sec}$, $\mathscr{L}_{id}$, and $\mathscr{L}_{rec}$ can produce a loss function which is considered as our regularization term. According to that,

$$\mathscr{L}_{reg} = \mathscr{L}_{sec} + \mathscr{L}_{id} + \mathscr{L}_{rec} \tag{14}$$

This is the general expression of our regularization term, which consists of several L_2 norm minimizations without any addition of fixed or variable hyperparameters. The regularization term forces the network directly to prevent the basic prerequisites of denoising, like avoiding over-smoothness with proper reconstruction.

Our total training objective function $\mathscr{L}_{total}$ is defined by the summation of all the aforementioned loss functions as follows:

$$\mathscr{L}_{total} = \mathscr{L}_1 + \mathscr{L}_{reg} = \mathscr{L}_1 + \mathscr{L}_{sec} + \mathscr{L}_{id} + \mathscr{L}_{rec} \tag{15}$$

V. Experiments

This section represents a discussion about the datasets and the implementation details for training and the evaluation procedure. The evaluation part describes the effectiveness of the proposed method through representing both qualitative and quantitative comparison with recent state-of-the-art methods for both 1) synthetic images with Gaussian and Poisson noise in sRGB space and 2) real-world noisy images in s-RGB space.

A. Training details.

Like other self-supervised methods, we propose a model-dependent noisy image reconstruction strategy. Our training dataset is the DIV2K dataset [51] consisting of 800 training images. During training, the patch extraction is applied with the size 40×40 on the 800 training images and augmented with only rotation due to anisotropy of depth. Hence the total training images are extended to 20000 images. The noisy version of all images is generated by adding Additive Gaussian Noise (AWGN) with specific noise levels. We employ two separate training procedures using DnCNN [5] and U-Net[2] as denoisers. For DnCNN, we employ the same baseline DnCNN model with the number of layers 17 to reconstruct the denoised output. For U-Net [2], we use the same modified architecture as [34]. Over the training procedure, each model is trained for 400 epochs with batch size 16 where all images are normalized between 0 to 1. The optimizer is Adam with a learning rate initialized to 0.0001. We utilize the cosine learning rate decay from the original Tensorflow library. All experiments are implemented under a server with Python 3.7, Tensorflow 2.2.0, and Nvidia GeForce GTX 3020 GPUs. After training, we save two weights of our method, one is for DnCNN, and another is for U-Net.

B. Synthetic noise removal experiments.

We test the denoising performance on the RGB version of three denoising datasets i.e. BSD68, Kodak24, and Set14. We compare our proposed method against one baseline method like Noise2Noise(N2N)[24], non-learning method CBM3D [11] for Gaussian and Anscombe [21] for Poisson, an unsupervised method Recorrupted2Recorrupted (R2R)[33], and four self-supervised methods: Noisier2Noise (Nr2N)[32], Noise2Self(N2S)[26] , Neighbor2neighbor (NBR2NBR) [34], and Blind2Unblind(B2UB) [31].

We follow the necessary modifications of the compared methods to represent a fair comparison with them. 1) For CBM3D [11], we follow the same procedure as [34] through using variance estimation for Gaussian and for Poisson to Gaussian noise conversion; 2) For Noise2Noise (N2N) [24], and Noisier2Noise (Nr2N) [32]; we re-implement the two methods using Keras-Tensorflow Library . We use the same DnCNN model of our method and train the model on the same amount of images of DIV2K dataset. Over the training time, the DnCNN is trained for 400 epochs with batch size 16, where all of the images are normalized between 0 to 1. After the training procedure, we test both of the methods on the above three test datasets using the pre-trained weights. 3) As both Noise2Self (N2S) and Recorrupted2Recorrupted (R2R) have used the DnCNN model like us, we use the pre-trained network weights provided by them and follow their own noise generation strategies. However, their official implementation is only for gray-scale images. We re-implemented their methods for the sRGB images to represent the visual results. 4) As Neighbor2Neighbor(NBR2NBR) and Blind2Unblind(B2UB) used U-Net architecture, we provide both visual and quantitative results according to their implementation. For fair comparison in

both sections, we provide our visual and quantitative results using DnCNN and U-Net architecture.

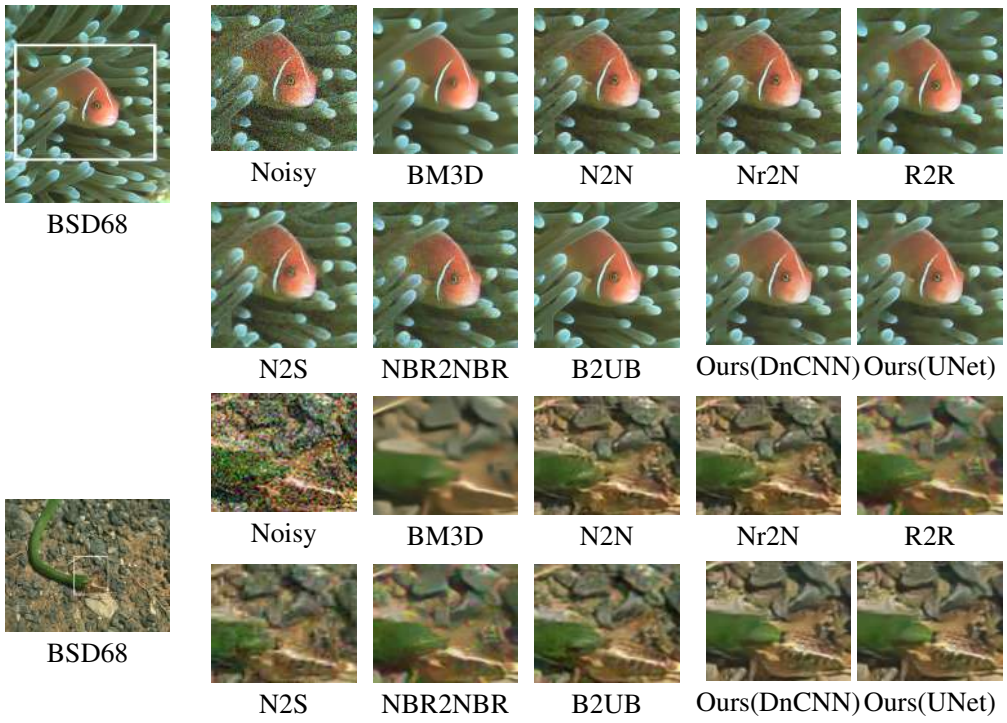

Figure 3: Visual comparison of denoising sRGB images of BSD68 recorrupted by AWGN $\sigma = 50$.

C. Results of Synthetic Experiments.

For the synthetic noise experiments, we provide both qualitative and quantitative comparison for two synthetic noise distributions between the methods mentioned above: 1) Gaussian noise (AWGN), and 2) Poisson noise. We provide at least two images from each dataset to show our effective visual performance. For

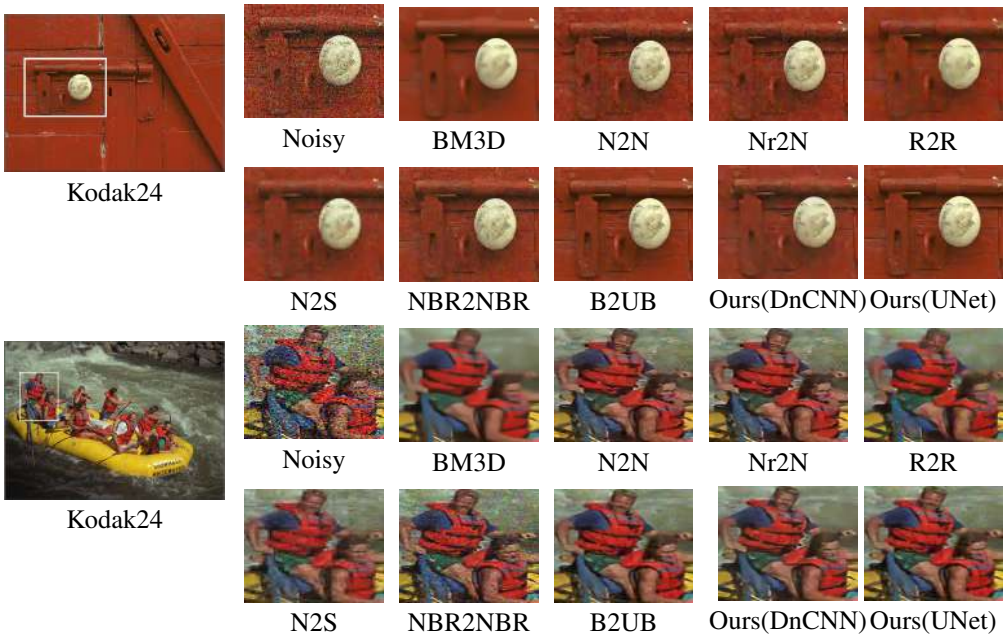

Figure 4: Visual comparison of denoising sRGB images of Kodak24 recorrupted by AWGN $\sigma = 50$.

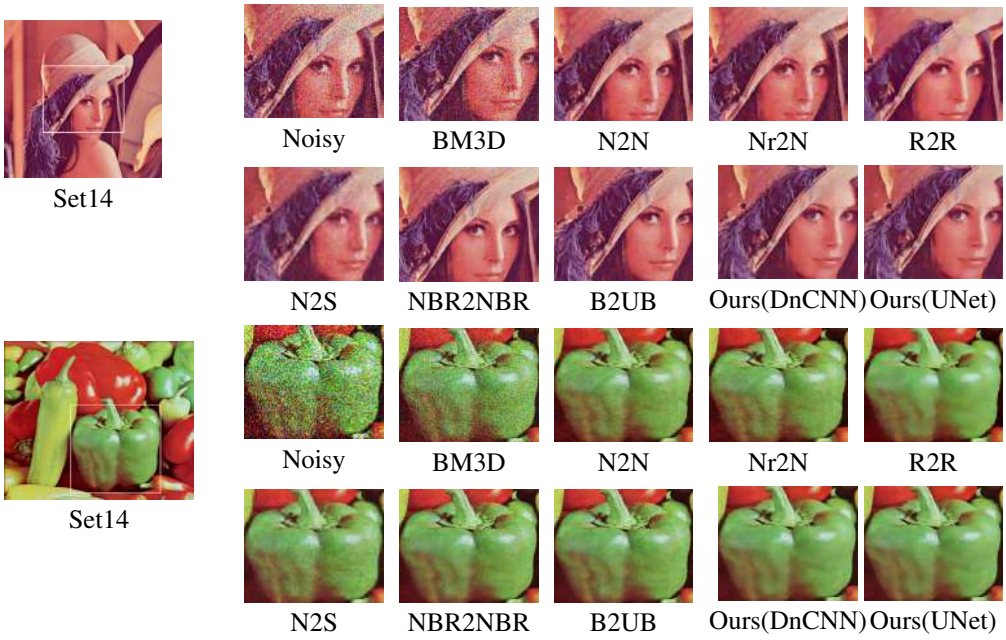

Figure 5: Visual comparison of denoising sRGB images of Set14 recorrupted by AWGN $\sigma = 50$.

Noise Type	Supervision	Method	Model	BSD68	Kodak24	Set14
$\sigma = 25$	Non-learning	CBM3D		28.56/0.801	31.87/0.868	30.88/0.854
	Supervised	DnCNN		29.19/0.830	30.02/0.907	31.11/0.878
	Weakly-supervised	N2N	DnCNN	31.01/0.866	32.41/0.884	31.37/0.868
	Un-supervised	R2R	DnCNN	30.81/0.822	29.98/0.906	31.32/0.865
	Self-supervised	Nr2N	DnCNN	28.55/0.808	29.39/0.893	29.64/0.832
		N2S	DnCNN	28.12/0.792	29.24/0.903	29.22/0.822
		NBR2NBR	U-Net	29.28/0.812	29.08/0.879	31.09/0.864
		B2UB	U-Net	30.89/0.875	32.27/0.880	**31.27/0.864**
		Proposed	DnCNN	**30.07/0.848**	**29.19/0.830**	**29.23/0.802**
		Proposed	U-Net	**33.70/0.881**	**33.00/0.891**	30.48/0.812
$\sigma = 50$	Non-learning	CBM3D		25.62/0.687	27.02/0.813	26.32/0.813
	Supervised	DnCNN		25.92/0.718	26.12/0.812	26.08/0.825
	Weakly-supervised	N2N	DnCNN	31.02/0.858	32.38/0.878	31.39/0.863
	Un-supervised	R2R	DnCNN	26.01/0.798	26.65/0.801	26.12/0.749
	Self-supervised	Nr2N	DnCNN	25.61/0.681	25.12/0.744	25.98/0.723
		N2S	DnCNN	25.88/0.792	26.24/0.903	26.22/ 0.814
		NBR2NBR	U-Net	26.13/0.709	27.12/0.849	26.03/0.813
		B2UB	U-Net	30.82/0.859	31.31/0.869	**31.08/0.849**
		Proposed	DnCNN	**26.46/0.814**	**27.49/0.852**	**26.18/ 0.812**
		Proposed	U-Net	**31.32/0.861**	**31.94/0.872**	28.69/0.856
$\lambda = [5,50]$	Non-learning	Anscombe		29.77/0.851	31.19/0.861	26.02/0.842
	Wealy-supervised	N2N	DnCNN	29.65/0.844	29.78/0.848	30.02/0.842
	Un-supervised	R2R	DnCNN	29.14/0.732	29.28/0.732	28.77/0.765
	Self-supervised	Nr2N	DnCNN	28.13/0.812	28.12/0.822	28.11/0.825
		N2S	DnCNN	28.93/0.823	28.08/0.808	27.62/0.835
		NBR2NBR	U-Net	30.86/0.855	29.54/0.843	29.79/0.838
		B2UB	U-Net	30.28/0.864	31.64/0.871	30.46/0.852
		Proposed	DnCNN	**31.18/0.814**	**27.49/0.852**	**28.82/ 0.812**
		Proposed	U-Net	**30.51/0.851**	**31.98/0.876**	**30.48/0.823**

Table 2: Quantitative comparison, in PSNR(dB)/SSIM, of different methods for AWGN removal on BSD68, Kodak24, and Set14. The compared methods are categorized according to the type of training samples

quantitative comparison, we provide the PSNR and SSIM results of the existing and our method on the above datasets.

1. Gaussian Noise removal result

We show both visual and quantitative performance comparisons under fixed Gaussian Noise. The following sections will provide a detailed comparative discussion about both visual and quantitative results between the proposed and the existing methods.

Visual Comparison: Fig. 3 and 4 illustrate visual denoising performance on the RGB versions of the BSD68 and Kodak24 datasets, which are corrupted by AWGN $\sigma = 50$. Our method achieves better recovery performance considering the image reconstruction performance during denoising. In the first image of Fig. 3, the challenge is reconstructing the original information about the fish and sea flora. We provide the BM3D result to show a comparison scenario between supervised and self-supervised methods. As a supervised method, BM3D reconstructs fine details from the noisy image. However, despite being a weakly supervised method, Noise2Noise(N2N) shows better visual performance than BM3D. The unsupervised method, Recorrupted2Recorrupted(R2R) shows an over-smooth output image. Among self-supervised ones, a mask-based blind method

like Noise2Self(N2S) suffers from over-smoothness with inconsistent structure recovery. On the other hand, unblind methods like Noisier2Noisier(Nr2N) and Neighbor2Neighbor(NBR2NBR) achieve satisfactory performance but still face either a blurry effect or noisiness. However, the very recent mask-based method, Blind2unblind(B2UB), performs better recovery of image details after

transitioning from blind to unblind spots. Similarly, our method efficiently reconstructs the image's fish and flora. The *"snake"* image from the BSD68 dataset is also challenging to reconstruct the complex background behind the snake. As the existing methods use both DnCNN and UNet model, we provide visual results using both denoisers for comparison. Our approach using both DnCNN and UNet reconstructs better through retrieving the original details of the complex background for both images in comparison to others.

In Fig. 4, the first image is the *"door"* image, where retrieving the original details of both the door and handle is challenging. Surprisingly, our method and Blind2unblind(B2UB) show better reconstruction than the supervised and weakly-supervised methods. On the other hand, the unsupervised method Recorrupted2Recorrupted (R2R), and Noise2Self(N2S) suffer from over-smoothness. Similarly, reconstructing both flows of waves and human faces is challenging in the second image. Moreover, the color contrast of the image is also an issue. The overall performance shows that our method achieves satisfactory visual results compared to others.

Fig. 5 represents two images from the dataset Set14. The first image called *"lena"* is challenging due to reconstructing the face and the background. Our method shows better denoising and reconstruction in comparison to others. The second image called *"Peppers"*, is another challenging image to retrieve the original details of the peppers. Here, we also show competitive performance. Blind2Unblind(B2UB) reconstructs better original information than ours. Our method suffers from over-smoothness like Recorrupted2Recorrupted (R2R) in this image. Interestingly, in some images from different datasets, mask-based methods perform better reconstruction while they generally suffer from information loss due to replacing the original information with masking.

Quantitative comparison: We also provide table 2 for a quantitative comparison between the above-mentioned methods. The Table represents the PSNR and SSIM on the three testing datasets corrupted with AWGN of two fixed noise levels $\sigma = 25, 50$ on the three test datasets. For BSD68, our method outperforms other existing methods in both fixed noise levels. Surprisingly, our method performs around 2db better than the supervised BM3D and the weakly supervised method Noise2Noise(N2N) with noisy/noisy image pairs. One possible cause might be that Noise2Noise(N2N) can only utilize the provided noisy pairs while our method can generalize multiple instances of image pairs from a single noisy image like Noisier2Noise(Nr2N) and Recorrupted2Recorrupted(R2R). Even our method achieves better performance than the above two methods under variable inherent noise, which amount is unknown. In comparison to the representative supervised method DnCNN, the performance gap between our and DnCNN is even better than Recorrupted2Recorrupted(R2R). We achieve around 1dB in PSNR in this case. For Kodak24 and Set14, our method achieves moderate or better performance compared to others in both noise-fixed Gaussian noise levels. We also compare our performance using U-Net architecture with the methods that have used U-Net architecture [31], [34]. For both BSD68, and Kodak 24, our method using U-Net achieves an increment of around 2db compared to the most recent method Blind2Unblind (B2UB). For Set14, our method using U-Net achieves the second best performance.

2. Poisson noise removal

We also show visual comparison under fixed Poisson Noise and variable quantitative comparison under variable Poisson Noise. Our effective performance under diverse noise distribution can prove the generalization performance of our method.

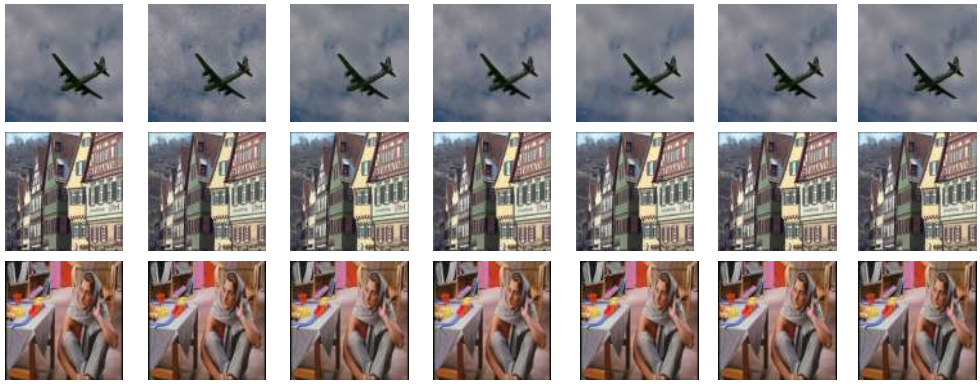

Figure 6: Visual comparison of the results from different methods when the denoising images recorrupted by Poisson $\lambda = 30$. The three images of the three columns are adopted from BSD68, Kodak24, and Set14 respectively.

Visual Comparison: For visual comparison, we show three individual images from the above three test datasets as shown in Fig. 6. For BSD68, we adopt the challenging *"aeroplane"* image for showing visual performance under fixed Poisson Noise $\lambda = 30$. The foreground-background difference of the image is completely contrastive in this image. As a result, the major challenge is retrieving the sky's cloud details. Moreover, retrieving the body design of this airplane is also crucial. According to the visual results, Recorrupted2Recorrupted(R2R) and

our method can perfectly retrieve the original details of both plane and sky. In contrast, the masked-based methods suffer from some undesired artifacts or over-smoothness problems due to replacing the original information with the random value of maskers.

Similarly, we show the visual comparison for the *"building"* image from the Kodak24 dataset for the same fixed Poisson Noise. The image is also challenging because an effective method should retrieve the spatial details like calligraphy, terracotta-bricks of the building. The mask-based methods Noise2self(N2S) drastically degrade their performance to recover the above spatial details. Even they fail to recover the bush of the background. In contrast, we as well as the unblind methods like Neighbor2neighbor(NBR2NBR) and Recorrupted2Recorrupted(R2R) have shown satisfactory performance in both foreground and background. If we analyze critically, we show even better performance in bush-detail recovery than the above unblind methods.

The last image is a case of maintaining the color contrast. In comparison to the mask-based methods, the unblind methods can maintain better color contrast. However, like other cases, Recorrupted2Recorrupted(R2R) suffers from over-smoothness. In contrast, our method recovers the detail avoiding over-smoothness effectively.

Quantitative Comparison: The last portion of the table 2 also shows variable Poisson Noise where the $\lambda = [5, 50]$. For variable noise experiments, the noise level λ is randomly generated between the range 5 to 50. For this comparison, we replace BM3D [11] with Anscombe transform [21] as the method is optimized only for the Poisson noise model. For variable Poisson Noise, our method performs better than three unblind and three mask-based methods.

However, for Kodak24, though our method shows moderate value performance, we achieve better reconstruction than the existing ones. For Set14, our method using U-Net achieves the best performance compared to the recent method Blind2Unblind(B2UB) [31].

To summarize the overall performance for both fixed and variable noise levels of AWGN and Poisson Noise, our method achieves a generalizable performance compared to others.

D. Real-noise removal experiments.

We perform our real noise removal experiments on four different datasets *i.e.* SIDD validation and benchmark [52], CC [53], and PolyU [54] dataset. Among these, only CC, PolyU, and SIDD Validation data provide the ground truth. Considering the priority of the datasets, we divide the real-noise experiments into two sections. The initial section evaluates SIDD Validation and Benchmarks in sRGB space. The second section is experimenting on CC and PolyU datasets. The corresponding website provides the evaluation result on SIDD Benchmark by submitting the denoised images. Both SIDD validation and Benchmark images are captured from 40 different scenes that are cropped into 32 blocks of size 256×256. The CC and PolyU datasets contain 15 and 100 RGB images respectively.

Experiments on real-world images(sRGB). Unlike the raw-RGB images, sRGB image values are standard across multiple camera devices. The sRGB images are processed through several camera-specific photo manipulations to make these visually pleasing. Since most of the digital images are stored as the sRGB format, any denoising method should perform robustly in sRGB space of

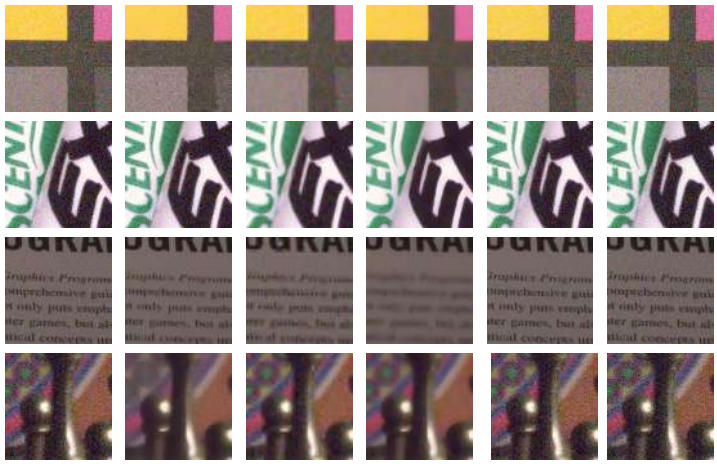

Figure 7: Visual comparison of the results from different methods when denoising an example image from dataset SIDD benchmark.

the real-world noisy images.

Datasets/Methods	CBM3D	NC	N2Self	Self2Self	NBR2NBR	Blind2unblind	DnCNN	R2R	Ours
SIDD Benchmark	25.65/0.685	31.26/0.826	29.56/0.808	33.38/0.846	34.01/0.858	34.08/0.812	**36.54/0.927**	34.78/0.898	34.12/0.866
SIDD Validation	25.65/0.475	31.31/0.725	30.72/0.787	35.04/0.902	34.68/0.782	34.62/0.822	**36.83/0.870**	35.04/0.844	34.98/0.858

Table 3: Quantitative comparison, in PSNR(dB)/SSIM, of different methods for denoising real-world images from SIDD.

For this experiment, we have followed the same training procedure of cvf-sid [55]. For training, we use the sRGB images from the SIDD Medium dataset while Recorrupted2Recorrupted(R2R) [33] requires the raw RGB images for pretraining to experiment on sRGB space. Moreover, raw RGB images contain more color information than the sRGB images, which commits better performance. As a result, we retrain the Recorrupted2Recorrupted(R2R) model on only sRGB images of the SIDD Medium dataset. For Noise2Self(N2S) [26],

we follow the same procedure for training. Self2Self(S2S) [36] is a single-shot denoising method. We follow the same procedure for this method which the authors provide. Our evaluation procedure is on both SIDD validation and SIDD Benchmark. We represent a quantitative comparison of this evaluation in table 3. For this comparison, we conduct blind image denoising method i.e. Noise Clinic (NC) [56] specifically designed for real-world images in addition to the non-learning method CBM3D [11] and deep learning methods i.e. Noise2Self(N2S)[26] and Recorrupted2Recorrupted (R2R) [33].

Result analysis of SIDD. Fig. 7 represents the visual comparison between our method and the existing methods on the SIDD benchmark. We provide some challenging images where texture reconstruction is crucial. In comparison to other methods, we perform better texture recovery during denoising. For quantitative comparison, table 3 represents a quantitative comparison on both SIDD validation and benchmark. Due to being a supervised model, DnCNN achieves the best performance. The unsupervised method Recorrupted2Recorrupted(R2R) [33] shows the second-best result. Our method shows competitive performance in comparison to the first and second-best methods.

E. Experiments on CC and PolyU.

CC dataset contains the real-world noisy images captured by Nikon D80 cameras. On the other hand, the PolyU dataset also contains real-world noisy images captured by different camera brands i.e. Nikon, Canon, and Sony through changing the ISO and shutter speed. As a result, the images of both datasets are much more complex than AWGN, signal-dependent, and depend on the diverse camera settings. Moreover, the noise of the images is inhomogeneous. For example, the overall noise standard deviations differ according to the color channels. Several methods [57] estimate the noise of different channels using noise level estimation. However, noise estimation for each individual channel achieves unsatisfactory performance with unwanted artifacts [58]. Remaining methods [59] concatenate patches of three channels into a single vector which is unable to consider diverse noise statistics among different channels. As a result, we apply the single shot denoising procedure for simplicity to learn the model about the noise distribution of CC and PolyU images without using any noise level estimator or patch concatenation.

CC and PolyU do not contain any training datasets, and the ground truth provided by both datasets is captured in low ISO settings and other post-processing like spatial alignment, varying intensity, low-frequency residual connection, etc.

So, we train the denoiser directly on both datasets' noisy images. To obtain the results of the DnCNN model for prediction, we use the pre-trained blind DnCNN model, which is trained over the color version of DIV2K with AWGN where the noise level is uniformly sampled from $[0,50]$. Then we set $\mathscr{M}_1 = \mathscr{M}_2$ with a very low noise amount because the noise level of CC and PolyU images is very low,

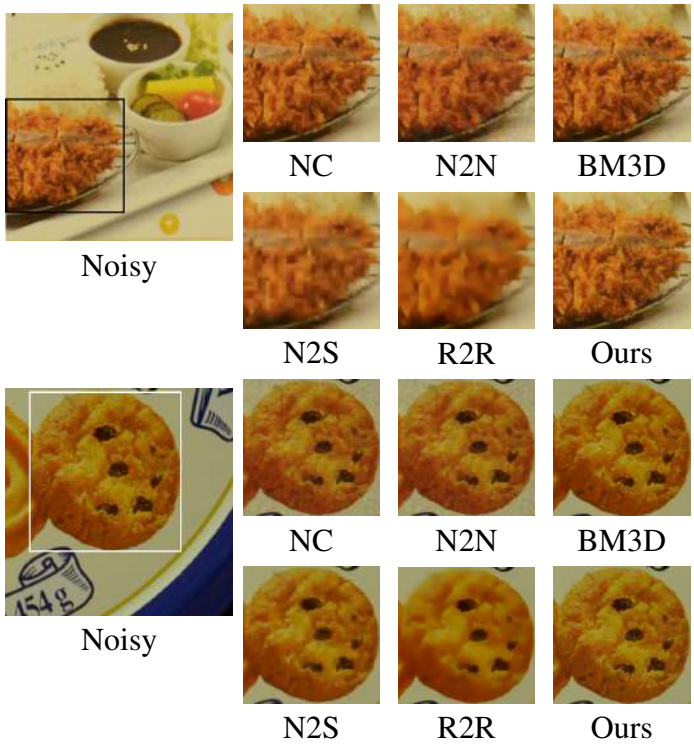

Figure 8: Visual comparison of the results from different methods when denoising an example image from dataset CC.

and heavier corruption is better for avoiding overfitting. For each image of size $512 \times 512 \times 3$, we train our DnCNN model with around 1000 iterations using a learning rate of 10^{-3}. It takes around half an hour to process a single image with size $512 \times 512 \times 3$ for our method.

For comparison, we provide a similar comparison scenario. We achieve a competitive performance considering other methods.

Visual Comparison of CC and PolyU: See Fig. 8 and 9 for some visual results. Our method can efficiently recover the original information of the image

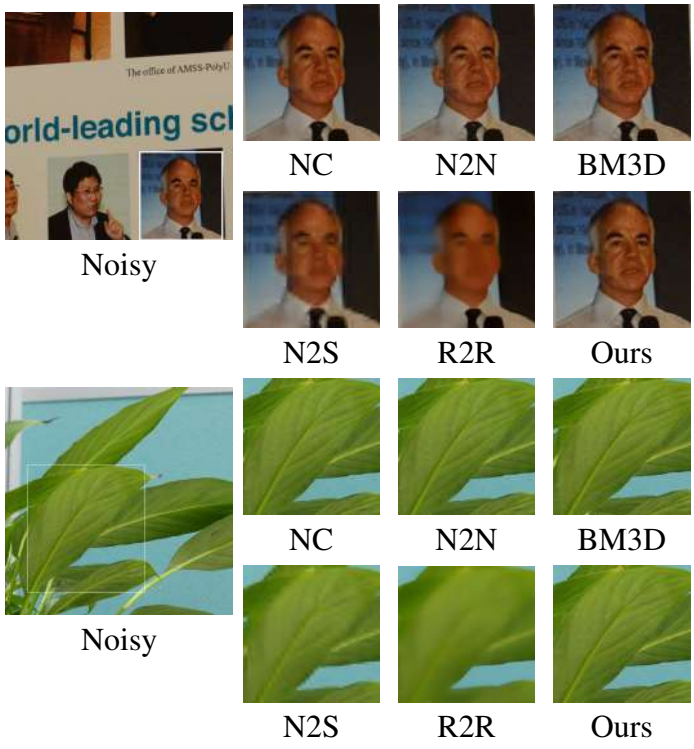

Figure 9: Visual comparison of the results from different methods when denoising an example image from dataset PolyU.

during denoising. For visual comparison, we adopt two images from each dataset, where the main focus is a reconstruction of the original detail. Both images of the CC dataset are "food" images, and our method outperforms another method through denoising and reconstruction. Through the PolyU visual results, we also show better performance in reconstructing the faces of the existing person and the details of the leaves. Overall, our method shows noticeable performance in both CC and PolyU datasets.

Quantitative Comparison of CC and PolyU: For these datasets, with the representative non-learning based methods CBM3D [11], we add two methods that are specifically designed for denoising, i.e., Multi-channel Weighted Nuclear Norm Minimization (MCWNNM) [12] and Noise Clinic (NC) [56]. Moreover, we add the DnCNN, N2V-single, N2S-single, R2R-single as an extension of Noise2Void(N2V)[25], Noise2Self(N2S) [26], Recorrupted2Recorrupted(R2R) [33] respectively. Among these methods, MCWNNM is exceptionally better and this method is sensitive to this kind of noise present in CC and PolyU datasets. Recorrupted2Recorrupted(R2R)[33] also achieves outstanding performance as they model the noise by AWGN with different noise levels for different color channels like MCWNNM [12]. Our method shows competitive performance with these two methods. Table 4 shows a comparison of both CC and PolyU datasets with several existing methods.

Datasets/Methods	CBM3D	MCWNNM	NC	N2V-single	N2S-single	S2S	DnCNN	R2R-single	Proposed
CC	35.19/0.858	37.70/0.954	33.38/0.846	33.47/0.932	35.64/0.859	36.81/0.913	33.47/0.932	**37.78/0.951**	37.46/0.899
PolyU	37.40/0.953	36.92/0.945	35.04/0.902	38.37/0.962	36.21/0.858	37.11/0.898	35.60/0.964	**38.47/0.965**	38.13/0.912

Table 4: Quantitative comparison, in PSNR(dB)/SSIM, of different methods for denoising real-world images from CC and PolyU

VI. Applications

Let us consider the outdoor images where the scenarios continuously change due to several reasons, like changes in sunlight or unfavorable weather conditions. Digital image-capturing systems can facilitate the capturing capability of stationary and moving objects with high-end digital cameras or mobile phones in such outdoor scenarios. However, CC/Surveillance cameras or real-life streaming scenarios still need help capturing images without any noise in different daylight conditions. As a result, the problems cause hindrances in several practical tasks like face detection, object detection, or person identification which is an emergency for such highly secured systems. Furthermore, the role of CC/Surveillance cameras is to capture images in a continuous process. So, such applications can't provide clean ground truths of the corresponding captured images. Therefore, a self-supervised denoising strategy can be a complete solution to denoise the images captured by the above systems.

Considering the issues above, we implement our proposed method for two practical applications: multi-face detection and object detection. The motivation behind the experiment is a proper reconstruction of the captured images without requiring any clean ground truths. Following such experiments, the highly restricted zones can develop the security system through the proper face and object detection and recognition.

A. Multiface detection

Our initial approach provides some experiments related to multi-face detection. For these experiments, we employed two renowned face detectors: RetinaFace [60], and MTCNN [61], which still perform the SOTA performance in the face detection domain. Like previous experiments, we recall a weakly-supervised

method Noise2Noise [24] and an unsupervised method R2R[33] to create a comparison scenario. We evaluate the methods on two multi-face datasets: AFW [62], and FDDB [63]. We produce the noisy images for both datasets with AWGN $\delta = 50$.

Figure 10: Experiments of multiface detection. The first two rows represent the output of MTCNN face detector on two individual images of AFW and FDDB datasets. The last two rows are the output of RetinaFace detector on the same images of the same datasets.

Fig. 10 represents the multiface images of three above-mentioned methods which are detected by MTCNN and RetinaFace detectors. Our intuition is that how much faces present in the denoised images can be detected by the face detectors in comparison to the clean images. Obviously, it is tough for the detectors to detect all of the faces in the noisy images. From that perspective,

the output images of the best denoising method will achieve the best detection performance. In comparison to the other two denoising methods, our method achieves the best detection performance through detecting all of the faces present in the images.

B. Object detection

Similarly, our secondary approach provides several experiments related to the object detection. For these experiments, we employed three object detection models You Only Look Once YoloV3 [64], YoloV5 [64], and YoloV6 [65]. Here we also recall the weakly supervised method Noise2Noise(N2N) [24], and the unsupervised method Recorrupted2Recorrupted(R2R) [33] to create the same comparison scenario. For object detection, we evaluate our methods on three datasets: CamVid [66], Kitty [67], and ECP [68]. For all datasets, we produce the noisy images by adding AWGN $\delta = 25$.

Fig. 11 represents the object detection images of three denoised methods, which YoloV3, YoloV4, and YoloV5 object detectors detect. From the figure, our method detects the most number of objects present in the images. Generally, the object detection datasets consist of some sequential images captured within very short elapsed times. As a result, it is quite impossible to detect moving objects like humans, cars, buses, trucks, etc., while the images are noisy. So, image denoising is very crucial in such practical applications. YoloV6 performs better than the previous versions based on the noisy image detection performance. While we compare the clean image detection with the denoised images detection of the three denoising methods, our method using the object detectors detects the highest number of moving objects in the image.

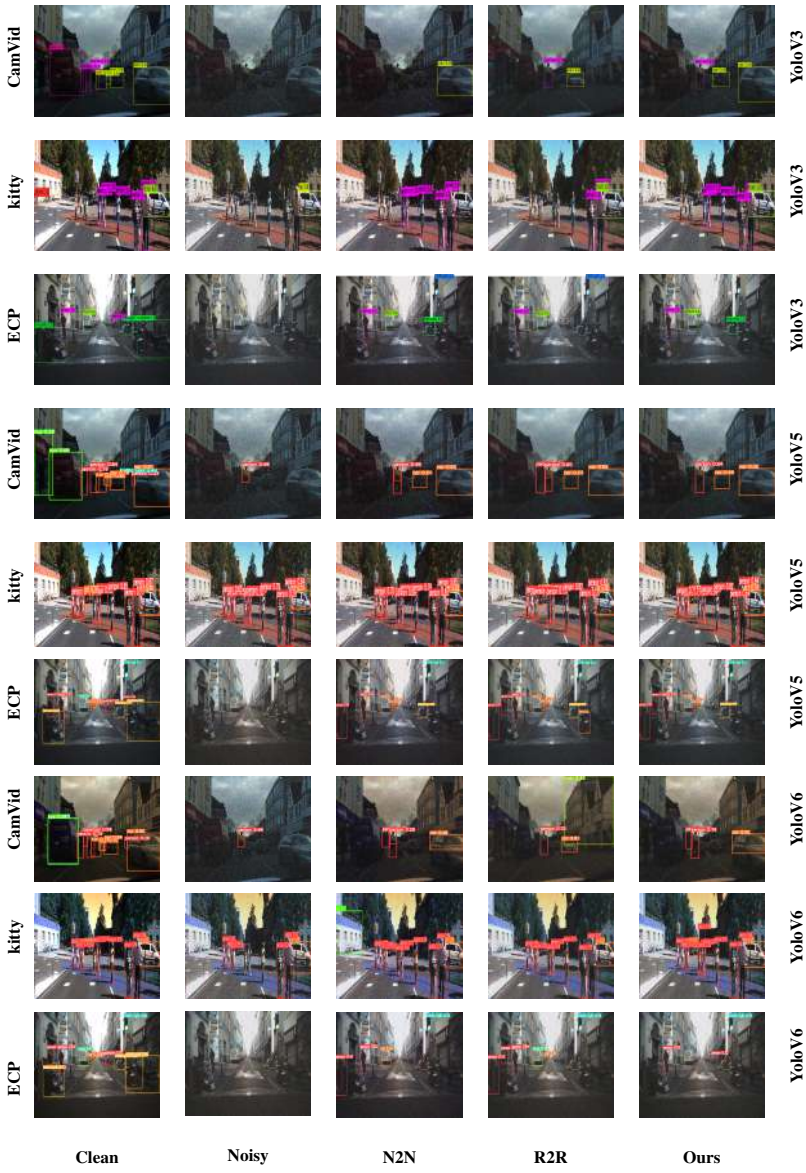

Figure 11: Experiments of object detection. The first three rows represent the output of the YoloV3 object detector on three individual images of CamVid, Kitty, and ECP datasets. The second rows are the outputs of the YoloV5 object detector on the same images of the same datasets. The last three rows are from the outputs of the YoloV6 object detector.

In summary, we provide these experiments to show how low-level vision tasks can improve high-level vision applications. In the long run, where collecting the ground truth images is so challenging, self-supervision can be an ultimate solution as denoising is crucial in these real-world applications.

VII. CONCLUSION

In summary, we enable the training scenario with complicated settings, which leads to a counter-intuitive approach in the case of noise recovery without any clean ground truth. Through the approach, we achieve competitive results between the existing ones. We demonstrate the inherent and additional noise uncertainty issue and establish a scenario to tackle this. However, the customized loss functions proposed for self-supervised methods are unable to ensure the convergence issue. Since self-supervised learning strategies use noisy labels instead of clean ground truth like supervised methods, it is complicated to guarantee the loss function's convergence. As a result, the question arises *what is the benefit of self-supervised learning in any vision domain?* The answer is that self-supervised strategy leads the way in designing any complicated loss function scenario to solve both classification and regression problems without requiring any clean ground truth. The expectation is the customized loss function can be able to fulfill the requirements of the above tasks.

To solve the denoising problem, our method follows a similar solution. Our customized loss function with additional regularization is able to produce noise-free images as far as possible with proper reconstruction and avoiding over-smoothness or any unwanted artifacts. Moreover, we represent some experiments for both synthetic and real noise domains by highlighting our effective performance. Finally, the implementation on different high-level vision applications can establish a scenario where alternate supervision improves the application performance while the ground truth is unavailable.

PUBLICATIONS

A. Journals

1. M. A. N. I. Fahim, N. Saqib, S. K. Siam *et al.*, "Rethinking gradient weight's influence over saliency map estimation," *Sensors*, **journvol** 22, **number** 17, **page** 6516, 2022.

2. M. A. N. I. Fahim, N. Saqib, S. K. Siam *et al.*, "Denoising single images by feature ensemble revisited," *Sensors*, **journvol** 22, **number** 18, **page** 7080, 2022.

3. M. A. N. I. Fahim, N. Saqib **and** J. H. Yub, "Semi-supervised atmospheric component learning in low-light image problem," *arXiv preprint arXiv:2204.07546*, 2022.

B. Conferences

1. N. Saqib **and** F. T. Zahra, "An improved adaptive optimization technique for image classification," **in***2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* IEEE, 2020, **pages** 1–6.

REFERENCES

[1] P. Goyal, M. Caron, B. Lefaudeux *et al.*, "Self-supervised pretraining of visual features in the wild," *arXiv preprint arXiv:2103.01988*, 2021.

[2] O. Ronneberger, P. Fischer **and** T. Brox, "U-net: Convolutional networks for biomedical image segmentation," **in***International Conference on Medical image computing and computer-assisted intervention* Springer, 2015, **pages** 234–241.

[3] B. Lim, S. Son, H. Kim, S. Nah **and** K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," **in***Proceedings of the IEEE conference on computer vision and pattern recognition workshops* 2017, **pages** 136–144.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He **and** P. Dollár, "Focal loss for dense object detection," **in***Proceedings of the IEEE international conference on computer vision* 2017, **pages** 2980–2988.

[5] K. Zhang, W. Zuo, Y. Chen, D. Meng **and** L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, **jourvol** 26, **number** 7, **pages** 3142–3155, 2017.

[6] Y. Chen **and** T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **jourvol** 39, **number** 6, **pages** 1256–1272, 2017. DOI: 10.1109/TPAMI.2016.2596743.

[7] X. Mao, C. Shen **and** Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," *Advances in neural information processing systems*, **journvol** 29, 2016.

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

[9] D. Liu, B. Wen, Y. Fan, C. C. Loy **and** T. S. Huang, "Non-local recurrent network for image restoration," *Advances in neural information processing systems*, **journvol** 31, 2018.

[10] T. Plötz **and** S. Roth, "Neural nearest neighbors networks," *Advances in Neural information processing systems*, **journvol** 31, 2018.

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

[12] J. Xu, L. Zhang, D. Zhang **and** X. Feng, "Multi-channel weighted nuclear norm minimization for real color image denoising," **in***Proceedings of the IEEE international conference on computer vision* 2017, **pages** 1096–1104.

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

[14] S. Lefkimmiatis, "Universal denoising networks: A novel cnn architecture for image denoising," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, **pages** 3204–3213.

[15] Y. Kim, J. W. Soh **and** N. I. Cho, "Adaptively tuning a convolutional neural network by gate process for image denoising," *IEEE Access*, **jourvol** 7, **pages** 63 447–63 456, 2019.

[16] C. Ren, X. He, C. Wang **and** Z. Zhao, "Adaptive consistency prior based deep network for image denoising," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2021, **pages** 8596–8606.

[17] S. Anwar **and** N. Barnes, "Real image denoising with feature attention," **in***Proceedings of the IEEE/CVF international conference on computer vision* 2019, **pages** 3155–3164.

[18] S. W. Zamir, A. Arora, S. Khan *et al.*, "Learning enriched features for real image restoration and enhancement," **in***ECCV* 2020.

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

[20] A. Foi, "Clipped noisy images: Heteroskedastic modeling and practical denoising," *Signal Processing*, **jourvol** 89, **number** 12, **pages** 2609–2629, 2009.

[21] M. Makitalo **and** A. Foi, "Optimal inversion of the generalized anscombe transformation for poisson-gaussian noise," *IEEE transactions on image processing*, **jourvol** 22, **number** 1, **pages** 91–103, 2012.

[22] S. Nam, Y. Hwang, Y. Matsushita **and** S. J. Kim, "A holistic approach to cross-channel image noise modeling and its application to image denoising," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2016, **pages** 1683–1691.

[23] K. Wei, Y. Fu, J. Yang **and** H. Huang, "A physics-based noise formation model for extreme low-light raw denoising," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2020, **pages** 2758–2767.

[24] J. Lehtinen, J. Munkberg, J. Hasselgren *et al.*, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[25] A. Krull, T.-O. Buchholz **and** F. Jug, "Noise2void-learning denoising from single noisy images," **in***Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, **pages** 2129–2137.

[26] J. Batson **and** L. Royer, "Noise2self: Blind denoising by self-supervision," **in***International Conference on Machine Learning* PMLR, 2019, **pages** 524–533.

[27] Y. Xie, Z. Wang **and** S. Ji, "Noise2Same: Optimizing a self-supervised bound for image denoising," **in***Advances in Neural Information Processing Systems* **volume** 33, 2020, **pages** 20 320–20 330.

[28] A. Krull, T. Vičar, M. Prakash, M. Lalit **and** F. Jug, "Probabilistic noise2void: Unsupervised content-aware denoising," *Frontiers in Computer Science*, **jourvol** 2, **page** 5, 2020.

[29] S. Laine, T. Karras, J. Lehtinen **and** T. Aila, "High-quality self-supervised deep image denoising," *Advances in Neural Information Processing Systems*, **jourvol** 32, 2019.

[30] X. Wu, M. Liu, Y. Cao, D. Ren **and** W. Zuo, "Unpaired learning of deep image denoising," **in***European conference on computer vision* Springer, 2020, **pages** 352–368.

[31] Z. Wang, J. Liu, G. Li **and** H. Han, "Blind2unblind: Self-supervised image denoising with visible blind spots," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2022, **pages** 2027–2036.

[32] N. Moran, D. Schmidt, Y. Zhong **and** P. Coady, "Noisier2noise: Learning to denoise from unpaired noisy data," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2020, **pages** 12 064–12 072.

[33] T. Pang, H. Zheng, Y. Quan **and** H. Ji, "Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2021, **pages** 2043–2052.

[34] T. Huang, S. Li, X. Jia, H. Lu **and** J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy images," **in***Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2021, **pages** 14 781–14 790.

[35] J. Xu, Y. Huang, M.-M. Cheng *et al.*, "Noisy-as-clean: Learning self-supervised denoising from corrupted image," *IEEE Transactions on Image Processing*, **journvol** 29, **pages** 9316–9329, 2020.

[36] Y. Quan, M. Chen, T. Pang **and** H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," **in***IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2020.

[37] S. Soltanayev **and** S. Y. Chun, "Training deep learning based denoisers without ground truth data," *Advances in neural information processing systems*, **journvol** 31, 2018.

[38] S. Ramani, T. Blu **and** M. Unser, "Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms," *IEEE Transactions on Image Processing*, **journvol** 17, **number** 9, **pages** 1540–1554, 2008, ISSN: 1941-0042. DOI: 10.1109/TIP.2008.2001404.

[39] T. Chen, K.-K. Ma **and** L.-H. Chen, "Tri-state median filter for image denoising," *IEEE Transactions on Image Processing*, **journvol** 8, **number** 12, **pages** 1834–1838, 1999. DOI: 10.1109/83.806630.

[40] J. Chen, J. Benesty, Y. Huang **and** S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, **journvol** 14, **number** 4, **pages** 1218–1234, 2006. DOI: 10.1109/TSA.2005.860851.

[41] S. Lefkimmiatis, "Universal denoising networks: A novel cnn architecture for image denoising," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, **pages** 3204–3213.

[42] Q. Xie, Q. Zhao, D. Meng *et al.*, "Multispectral images denoising by intrinsic tensor sparsity regularization," **in***2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016, **pages** 1692–1700. DOI: 10.1109/CVPR.2016.187.

[43] M. Elad **and** M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, **jourvol** 15, **number** 12, **pages** 3736–3745, 2006. DOI: 10.1109/TIP.2006.881969.

[44] S. Gu, L. Zhang, W. Zuo **and** X. Feng, "Weighted nuclear norm minimization with application to image denoising," **in***2014 IEEE Conference on Computer Vision and Pattern Recognition* 2014, **pages** 2862–2869. DOI: 10.1109/CVPR.2014.366.

[45] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang **and** L. Zhang, "Weighted schatten p-norm minimization for image denoising and background subtraction," *IEEE Transactions on Image Processing*, **jourvol** 25, **august** 2016. DOI: 10.1109/TIP.2016.2599290.

[46] J. Xu, L. Zhang, D. Zhang **and** X. Feng, "Multi-channel weighted nuclear norm minimization for real color image denoising," **in***Proceedings of the IEEE international conference on computer vision* 2017, **pages** 1096–1104.

[47] T. Xie, S. Li **and** B. Sun, "Hyperspectral images denoising via nonconvex regularized low-rank and sparse matrix decomposition," *IEEE Transactions on Image Processing*, **jourvol** 29, **pages** 44–56, 2020. DOI: 10.1109/TIP.2019.2926736.

[48] S. Lefkimmiatis, "Universal denoising networks: A novel cnn architecture for image denoising," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, **pages** 3204–3213.

[49] S. Cha, T. Park **and** T. Moon, "Gan2gan: Generative noise learning for blind image denoising with single noisy images," *arXiv preprint arXiv:1905.10488*, **jourvol** 3, 2019.

[50] J. Chen, J. Chen, H. Chao **and** M. Yang, "Image blind denoising with generative adversarial network based noise modeling," **in***Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, **pages** 3155–3164.

[51] E. Agustsson **and** R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," **in***The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* 2017.

[52] A. Abdelhamed, S. Lin **and** M. S. Brown, "A high-quality denoising dataset for smartphone cameras," **in***Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, **pages** 1692–1700.

[53] S. Nam, Y. Hwang, Y. Matsushita **and** S. J. Kim, "A holistic approach to cross-channel image noise modeling and its application to image denoising," **in***2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016, **pages** 1683–1691. DOI: 10.1109/CVPR.2016.186.

[54] J. Xu, H. Li, Z. Liang, D. Zhang **and** L. Zhang, "Real-world noisy image denoising: A new benchmark," *arXiv preprint arXiv:1804.02603*, 2018.

[55] R. Neshatavar, M. Yavartanoo, S. Son **and** K. M. Lee, "Cvf-sid: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image," **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2022, **pages** 17 583–17 591.

[56] M. Lebrun, M. Colom **and** J.-M. Morel, "The Noise Clinic: a Blind Image Denoising Algorithm," *Image Processing On Line*, **jourvol** 5, **pages** 1–54, 2015, https://doi.org/10.5201/ipol.2015.125.

[57] C. Liu, R. Szeliski, S. Bing Kang, C. L. Zitnick **and** W. T. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **jourvol** 30, **number** 2, **pages** 299–314, 2008. DOI: 10.1109/TPAMI.2007.1176.

[58] J. Mairal, M. Elad **and** G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, **jourvol** 17, **number** 1, **pages** 53–69, 2008. DOI: 10.1109/TIP.2007.911828.

[59] M. Lebrun, M. Colom **and** J.-M. Morel, "Multiscale image blind denoising," *IEEE Transactions on Image Processing*, **jourvol** 24, **number** 10, **pages** 3149–3161, 2015. DOI: 10.1109/TIP.2015.2439041.

[60] J. Deng, J. Guo, E. Ververas, I. Kotsia **and** S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," **in***Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2020, **pages** 5203–5212.

[61] K. Zhang, Z. Zhang, Z. Li **and** Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, **jourvol** 23, **number** 10, **pages** 1499–1503, 2016.

[62] X. Zhu **and** D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," **in***2012 IEEE Conference on Computer Vision and Pattern Recognition* 2012, **pages** 2879–2886. DOI: 10.1109/CVPR.2012.6248014.

[63] V. Jain **and** E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[64] J. Redmon **and** A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[65] C. Li, L. Li, H. Jiang *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[66] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez **and** J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[67] A. Geiger, P. Lenz **and** R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," **in***Conference on Computer Vision and Pattern Recognition (CVPR)* 2012.

[68] M. Braun, S. Krebs, F. Flohr **and** D. M. Gavrila, "The eurocity persons dataset: A novel benchmark for object detection," *arXiv preprint arXiv:1805.07193*, 2018.

ACKNOWLEDGEMENTS

My gratefulness begins with ALLAH, the creator of the whole universe and the negligible me. Then, I would like to acknowledge and thank my honorable supervisor, Professor Ho Yub Jung. His continuous guidance and support through providing adequate hardware resources and immense patience due to my foibles help me to find the such destination. Secondly, I am very grateful to my informal supervisor, my close friend Masud An Nur Islam Fahim. I have no words to explain my gratitude towards him. Getting an instructor and friend like him is a beautiful chapter of my life. I also give thanks to my labmate, my younger brother, Shafkat Khan Siam, for his immense support throughout my whole master's life.

Now the time is for taking a deep breath since I want to remember the memory of my prematurely dead father, a heartbreaking memory of my life. He was my life's pioneer, instructor, and superman. Secondly, I want to remember my beloved mother, whose continuous and selfless support is beyond description. Finally, my beloved wife, whose endless inspiration and belief in me, kept my spirits and motivation high during the process. Last of all, I want to give an ocean of love to my two siblings, Saba and Safa. ALHAMDULILLAH.