IDEA-Net: Adaptive Dual Self-Attention Network for Single Image Denoising

Zheming Zuo[1], Xinyu Chen[2], Han Xu[3,4], Jie Li[5] , Wenjuan Liao[6], Zhi-Xin Yang[2], Shizheng Wang[4]
[1]Department of Computer Science, Durham University, UK
[2]Department of Electromechanical Engineering, University of Macau, China
[3]School of Microelectronics, University of Chinese Academy of Sciences, China
[4]Institute of Microelectronics, Chinese Academy of Sciences, China
[5]School of Computing, Engineering & Digital Technologies, Teesside University, UK
[6]College of Engineering and Computer Science, Australian National University, Australia
zheming.zuo@durham.ac.uk, mb95408@connect.um.edu.mo, ann.gong.qifeng@gmail.com,
jie.li@tees.ac.uk, wenjuan.liao@outlook.com, zxyang@um.edu.mo, shizheng.wang@foxmail.com

Abstract

*Image denoising is a challenging task due to possible data bias and prediction variance. Existing approaches usually suffer from high computational cost. In this work, we propose an unsupervised image denoiser, dubbed as adapt**I**ve **D**ual s**E**lf-**A**ttention **Net**work (**IDEA-Net**), to handle these challenges. IDEA-Net benefits from a generatively learned image-wise dual self-attention region where the denoising process is enforced. Besides, IDEA-Net is not only robust to possible data bias but also helpful to reduce the prediction variance by applying a simplified encoder-decoder with Poisson* `dropout` *operations on a single noisy image merely. The proposed IDEA-Net demonstrated the outperformance on four benchmark datasets compared with other single-image-based learning and non-learning image denoisers. IDEA-Net also shows an appropriate choice to remove real-world noise in low-light and noisy scenes, which in turn, contribute to more accurate dark face detection. The source code is available at* `https://github.com/zhemingzuo/IDEA-Net`.

1. Introduction

Image denoising is arguably one of the most prevalent problems within the realms of image processing and computer vision [3, 4]. It aims to remove measurement noises or distortions from noisy images [47]. Fundamentally, image denoising could be treated as a process of leveraging the data bias [23, 41] and prediction variance [3, 11]. To cope with data bias, denoisers trained based on a single noisy image tends to be more robust in comparison to those trained on the entire external dataset [3, 13, 16]. Nevertheless, denoisers trained on a single noisy image are suffering two major challenges: 1) launching self-supervised learning efficiently with the lack of ground truth, 2) avoiding the prediction variance reduction [26]. Additionally, existing denoisers usually demand high computational cost, especially with respect to time complexity [43].

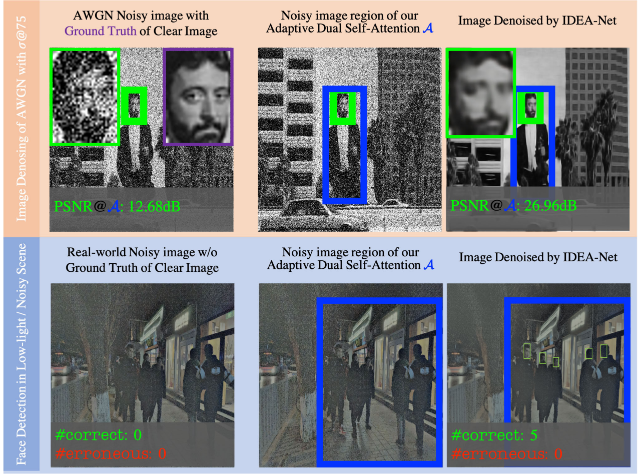
Image Denoising of AWGN with σ@75
AWGN Noisy image with Ground Truth of Clear Image
Noisy image region of our Adaptive Dual Self-Attention $\mathcal{A}$
Image Denoised by IDEA-Net
PSNR@$\mathcal{A}$: 12.68dB
PSNR@$\mathcal{A}$: 26.96dB
Face Detection in Low-light / Noisy Scene
Real-world Noisy image w/o Ground Truth of Clear Image
Noisy image region of our Adaptive Dual Self-Attention $\mathcal{A}$
Image Denoised by IDEA-Net
#correct: 0
#erroneous: 0
#correct: 5
#erroneous: 0

Figure 1. Performance of a versatile image denoiser could be revealed by not only a full-reference quality metric *e.g.* PSNR but also precision in a real-world task *e.g.* dark/noisy face detection.

Although several methods have successfully incorporated into image denoising, *e.g.* (C)DnCNN [44], FDnCNN [17] and Noise2Self [3], they are usually time-consuming, thus it still lacks of the computationally efficient denoising solution in the literature. In image denoising, given a clear image $\mathcal{I} \in \mathbb{R}^{W\times H\times L}$, the additive noise-corrupted image $\widetilde{\mathcal{I}} \in \mathbb{R}^{W\times H\times L}$ is constructed by

$$\widetilde{\mathcal{I}} = \mathcal{I} \oplus \mathbf{N}, \tag{1}$$

where $\mathbf{N}$ denotes the white Gaussian s.t. $\mathbb{E}(\mathbf{N}) = \mathbf{0}$ and

$\mathrm{Cov}(\mathbf{N}) = \sigma^2 \mathbf{I}_{W \times H \times L}$, $\oplus$ denotes element-wise addition. Motivated by the concepts of the potential of attention [32] and Region of Interest (RoI) [48], this paper devises a simple yet efficient deep-network-based image denoiser with versatility (see Figure 1), in which the training process requires an end-to-end learned dual-self attention region $\mathcal{A} = \{A_1, A_2\} \in \mathbb{R}^{R \times C \times L}$ within a single noisy image merely. Equivalently, this paper studies how to train a region-based image denoiser

$$\mathcal{M}_{\phi,\xi}: \quad \widetilde{\mathcal{I}_{\mathcal{A}}} \to \mathcal{I}_{\mathcal{A}}, \tag{2}$$

where $\mathcal{I}_{\mathcal{A}} \in \mathbb{R}^{R \times C \times L}$ and $\widetilde{\mathcal{I}_{\mathcal{A}}} \in \mathbb{R}^{R \times C \times L}$ respectively represents the clear and noisy image with image-wise dual-self attention $\mathcal{A}$ applied, and $RC \leqslant WH$. One step further, A_1 and A_2 respectively represents the first and second self-attention that learned via a variant of the unsupervised Cycle Generative Adversarial Network (CycleGAN) [22], denoted as $\mathcal{M}_\phi$ and termed as Dual Self-Attention Generative Adversarial Module (DSA-GAM). Built upon the dual-self attention region $\mathcal{A}$ introduced by $\mathcal{M}_\phi$, we design a simplified Encoder-Decoder Module (DSA-EDM) $\mathcal{M}_\xi$ for region-based image denoising in which `dropout` [28] operations are performed in line with Poisson distributions due to their benefits from preventing over-fitting and reducing prediction variance [10]. The objective of IDEA-Net is to minimise

$$\mathcal{L} = \operatorname*{argmin}_{\phi,\xi} \sum_k \mathtt{dist}\Big(\mathcal{M}_{\phi,\xi}(\widehat{\mathcal{I}_{\mathcal{A}}}), \widetilde{\mathcal{I}_{\mathcal{A}}} - \widehat{\mathcal{I}_{\mathcal{A}}}\Big), \tag{3}$$

where $\mathtt{dist}(\cdot,\cdot)$ represents the distance between the two images, $\mathcal{M}_{\phi,\xi}(\widehat{\mathcal{I}_{\mathcal{A}}})$ represent the denoising result yielded by the proposed IDEA-Net, and $\widetilde{\mathcal{I}_{\mathcal{A}}} - \widehat{\mathcal{I}_{\mathcal{A}}}$ denotes the unsampled part of $\mathcal{A}$ in the input noisy image. Furthermore, the size of $\mathcal{A}$ is varying for each input noisy image, thereby our denoiser is equipped with adaptive dual self-attention.

Combining DSA-GAM and DSA-EDM, we present a self-supervised attention network, termed as adaptIve Dual sElf-Attention Network (IDEA-Net), for single image denoising, and our contributions are summarised as follows:

1) We propose a self-supervised deep image denoiser merely requiring a dual self-attention region within a single input noisy image to appropriately handle possible data bias and dramatically reduce the computational cost.

2) We prove that a simplified encoder-decoder with Poisson `dropout` strategy could be better informed by the learned dual self-attention region in an ensemble learning manner to reduce the prediction variance.

3) We show that the proposed denoising scheme significantly outperforms the existing state-of-the-art methods for real-world face detection in low-light and noisy scenes.

Though existing denoisers either trained on a single input image or an additional dataset are utilising all the pixels within each training and/or testing instance, we prove that the region-based method could achieve better performance in solving traditional image denoising problem. In addition, we show that PSNR may not practically sufficient to distinguish the performance of denoising methods on real-world down-stream tasks such as dark face detection. The observations obtained from this work could also encourage more promising future work in pervasive healthcare such as medical image denoising [9, 38].

2. Related Work

Since deep learning methods are incorporated into image denoising, traditional learning-based methods are trained in a supervised manner using a set of clean and noisy image pairs. Practically, as ground truth is hard to acquire, learning-based image denoisers mainly use two different training strategies coping with only noisy images, learning from a single noisy image or a set of noisy images.

Denoisers learned from a single noisy image. In the early stage, taking advantage of the self-similarity, dictionary-based learning methods employ patches from the noisy image for training. As the icon, KSVD algorithm [8] is proposed to obtain trained dictionaries, which effectively describe the image content, with patches from the corrupted image. NCSR [7] learns the sub-dictionaries from the noisy image itself instead of the example clean images to get a more stable and sparser representation. A replacement of KSVD [2] designs a fast orthogonal dictionary learning method for decreasing the redundancy of the dictionary. TWSC [36] also utilises the KSVD dictionary learning scheme and introduces three weight matrices to characterise the statistics of realistic noise and image priors. Besides, with another form of self-similarity, the non-local approach (C)BM3D [5] increases its robustness using stacks of similar patches of the input noisy image and performs thresholding in frequency space. Thus, several methods [19, 6] combine the dictionary-based learning methods and the non-local approaches for better performance. However, it is noteworthy that Deep Image Prior (DIP) [31] is the pioneer work within the realm of single image denoising, which inspired the propositions of a series of methods including Noise2Noise [16], Noise2Self [3], Noise2Void [13], Self2Self (S2S) [26], etc. Therein a self-supervised learning method, S2S, is proposed to train the input noisy image merely with dropout on the pairs of Bernoulli-sampled instances and achieve remarkable performance enhancement.

Denoisers learned from a set of noisy images. Driven by the easy access to a large-scale dataset, the convolutional neural networks are training on a set of noisy images for tackling various vision tasks. For instance, the model-based optimisation methods can flexibly address different inverse tasks yet with high time complexity. (C)DnCNN

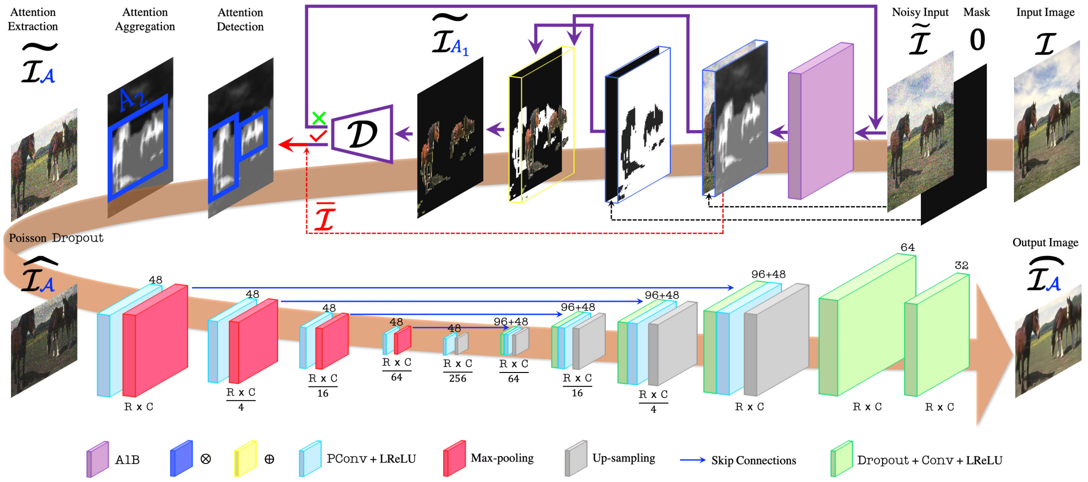

Figure 2. Architecture of the proposed IDEA-Net. Upper row present the DSA-GAM and lower row denotes the DSA-EDM.

[44] combines residual learning and batch normalisation to facilitate the training process and the denoising performance. While discriminative learning methods have fast testing speed but are limited within a specific task. FDnCNN [17] introduces a module named fusion block in CNNs to obtain high-quality images in real-time. Inheriting from DIP, Noise2Noise [16] employs pairs of noisy images with the same content for supervised learning, while Noise2Void [13] is a self-supervised training method with any noisy images. Then, Noise2Self [3] provides strong theoretical guarantees that Noise2Void has not by proposing $\mathcal{J}$-invariant. On the basis of self-supervised learning, a series of new methods are emerging. Noise2Inpaint [37] introduces a regularised image inpainting framework, Noise2Kernel [14] proposes a dilated convolutional network using kernel-based training, and Noise2Sim [24] utilises non-local mean to leverage self-similarities of image patches in self-learning. Therefore, integrating model-based optimisation and discriminative learning methods is a good attempt for further improvement. IRCNN [45] has trained a set of CNN denoisers and incorporate them into model-based optimisation method to maintain the good performance in various applications. ISCL [15] combines cyclic adversarial learning with self-supervised residual learning to boost the performance via cooperative learning.

3. IDEA-Net

In this section, we present the proposed IDEA-Net in detail. Briefly, we first explain the IDEA-Net architecture. This is followed by the introductions of the training and denoising schemes.

3.1. Architecture

The architecture of the proposed IDEA-Net is depicted in Figure 2. Briefly, it contains two consecutive modules: a Dual Self-Attention Generative Adversarial Module (DSA-GAM) and its associated simplified Encoder-Decoder Module (DSA-EDM). Specifically, DSA-GAM is proposed to generate an appropriate single attention region in the input noisy image, while DSA-EDM is devised to conduct denoising within the learned single attention region.

Given an input noisy image $\widetilde{\mathcal{I}} \in \mathbb{R}^{W \times H \times L}$ and a blackout mask $\mathbf{0} \in \mathbb{R}^{W \times H \times L}$ (*i.e.* image in the target domain), DSA-GAM feeds the noisy image into a subnet of five convolution layers, namely 1st Attention Block (A1B), to generate the first candidate attention $\bar{\mathcal{I}}$. The output of the generator in DSA-GAM is yielded by applying Eq. (5), which is checked by the discriminator $\mathcal{D}$ and its associated Least Square loss. Such a loss is adopted to force DSA-GAM to focus on the foreground attention region (*i.e.* 'real') rather than the background one (*i.e.* 'fake'). Once DSA-GAM reaches the maximum number of training iterations, the first candidate attention image $\bar{\mathcal{I}}$ is consequently fed into the consecutive 2nd Attention Block (A2B). A2B consists of three parts: detection, aggregation and extraction. In the detection stage, all the possible bounding boxes bbox are annotated surrounding each of the attentional sub-regions. Lastly, all the annotated bbox are aggregated into one single bbox (A_2), which in turn, extracts the attentional region $\widetilde{\mathcal{I}_{\mathcal{A}}} \in \mathbb{R}^{R \times C \times L}$ from $\widetilde{\mathcal{I}}$.

DSA-EDM first obtains $\widehat{\mathcal{I}_{\mathcal{A}}}$ by performing the Poisson dropout operation on $\widetilde{\mathcal{I}_{\mathcal{A}}}$. Then the sampled image $\widehat{\mathcal{I}_{\mathcal{A}}}$ is mapped to a $R \times C \times 48$ feature tube, which connects with an encoder. The DSA-EDM encoder contains 5 blocks, each of the first four blocks includes a Partial Convolution (PConv), a Leaky Rectified Linear Unit (LReLU), and a max-pooling operation with the stride of 2 and 2-by-2 receptive fields. The fifth encoder block only contains PConv and LReLU. Consequently, encoder results in a $R/16$-by-$C/16$-by-48 pixels feature cube when fixing the number of channels to 48 throughout all DSA-EDM encoder blocks. The DSA-EDM decoder contains 4 blocks and each of the first three blocks includes an up-sampling operation with a factor of 2 (denoted as `up-2`), a concatenation operation, a convolution with `dropout`, and LReLU. The first three decoder blocks have '`96+48`' channels in which '`48`' is contributed by its corresponding encoder block via the '`skip connection`'. In the last DSA-EDM decoder block, 3 convolution layers with LReLUs activation functions are adopted to map the feature cube back to the resolution of $\widetilde{\mathcal{I}_{\mathcal{A}}}$ (*i.e.* $R \times C \times L$). Thereby, the number of resulted channels of those convolution layers are respectively 64, 32 and L.

The architecture of our IDEA-Net shares similarity with the ones utilised in some of the existing methods *e.g.* S2S [26]. The key differences between IDEA-Net and S2S are three-fold. Firstly, we introduce the block `A1B` within DSA-GAM in which we design an all-black image as the target domain image (mask) to suit the problem domain of image denoising rather than image translation [22]. Secondly, we propose the block `A2B` in DSA-GAM to further process the output of `A1B` and generate the final dual self-attention region $\mathcal{A}$ to contribute the reduction of computational cost for the consecutive denoising module DSA-EDM. Lastly, we deploy DSA-EDM with higher Poisson sampling probability and shallower encoder/decoder blocks to make a tradeoff between time complexity (particularly with respect to the model convergence speed) and performance gain.

3.2. Training Scheme

As problem formulated in Eq. (2), we propose IDEA-Net by developing two consecutive modules DSA-GAM ($\mathcal{M}_\phi$) and DSA-EDM ($\mathcal{M}_\xi$). $\mathcal{M}_\phi$ is a cycle-generative module to learn an end-to-end dual self-attention region $\mathcal{A}$ from the input noisy image, which in turn, inform the process of region-based image denoising via an encoder-decoder module $\mathcal{M}_\xi$ integrated with Poisson `dropout` $\boldsymbol{p}$ strategy.

To obtain the first attention A_1, the cycle-generative module $\mathcal{M}_\phi$ is formulated as a special case of attention-based CycleGAN in which the target domain image $\widetilde{\mathcal{I}_Y} \in \mathbb{R}^{W\times H\times L}$ is set to be $\mathbf{0}$, the output image of the generator is generated in process yet eliminated in operation as it is not useful for image denoising, and expressed as

$$\mathcal{M}_\phi = \mathcal{M}_{Y\to X}\left(\mathcal{M}_{X\to Y}\left(\widetilde{\mathcal{I}_X}\right)\right) \approx \widetilde{\mathcal{I}_X}, \tag{4}$$

where X and Y respectively denotes the source and target image domain, $\widetilde{\mathcal{I}_X} := \widetilde{\mathcal{I}}$, and $\mathcal{M}_{X\to Y}(\widetilde{\mathcal{I}_X}) = \widetilde{\mathcal{I}_Y}$.

Practically, A_1 could be treated as a combination of foreground and background attention, computed as

$$\widetilde{\mathcal{I}_{A_1}} = \left(\overline{\mathcal{I}_X} \odot \widetilde{\mathcal{I}_X}\right) \oplus \left(\left(\mathbf{1} - \overline{\mathcal{I}_X}\right) \odot \mathbf{0}\right), \tag{5}$$

where $\widetilde{\mathcal{I}_{A_1}} \in \mathbb{R}^{W\times H\times L}$, $\odot$ represents the element-wise multiplication, $\overline{\mathcal{I}_X} := \texttt{A1B}(\widetilde{\mathcal{I}_X})$ where `A1B` denotes the 1st Attention Block. The `A1B` contains 5 convolution layers with residual and up-sampling operations in-between (detailed in Section 4.3).

Following the design as of [22], we implement the discriminator $\mathcal{D}$ that consists of four convolution layers and each of which contains zero padding and Leaky ReLU (LReLU). Given $\widetilde{\mathcal{I}_{A_1}}$ and $\widetilde{\mathcal{I}}$, DSA-GAM feeds the outputs of the discriminator $\mathcal{D}$ (*i.e.* foreground feature maps) into the Least Square loss [20] for minimisation as it helps generate sharper images.

When the maximum number of training iterations reached, $\mathcal{M}_\phi$ continues to conduct the 2nd Attention Block (`A2B`). `A2B` (as summarised in Algorithm 1) is devised to generate the final attention region A_2 and its associated $\widetilde{\mathcal{I}_{\mathcal{A}}}$, which includes attention detection (line 1-2), attention aggregation (line 3-9), and attention extraction (line 10).

Algorithm 1 2nd Attention Block (`A2B`)

```
Input: $\widetilde{\mathcal{I}}, \bar{\mathcal{I}}$, binarisation threshold $b$, area threshold $s$
Output: $\widetilde{\mathcal{I}_{\mathcal{A}}}$
 1: Compute the binarised image $\overline{\mathcal{I}_{\mathbf{b}}}$ w.r.t. $b$
 2: Detect and count the #contours $c$ in $\overline{\mathcal{I}_{\mathbf{b}}}$
 3: for $i = 1$ to $c$ do
 4:     Draw $\texttt{bbox}_i := \{x_i, y_i, w_i, h_i\}$ for each contour
 5:     if $w_i h_i > s$ then
 6:         $\{x_l \leftarrow x_i, x_r \leftarrow x_i + w_i, y_l \leftarrow y_i, y_r \leftarrow y_i + h_i\}_j$
 7:     end if
 8: end for
 9: $A_2 = \text{minmax}\{x_l, x_r, y_l, y_r\}_j$
10: return $\widetilde{\mathcal{I}_{\mathcal{A}}} = \widetilde{\mathcal{I}}[A_2]$
```

Since our IDEA-Net is trained on a dual self-attention region within a single noisy image $\widetilde{\mathcal{I}_{\mathcal{A}}}$, thus module $\mathcal{M}_\xi$ generates multi-pair of information-preserving images $\{(\widehat{\mathcal{I}^u_{\mathcal{A}}}, \widecheck{\mathcal{I}^u_{\mathcal{A}}})\}_{u=1}^U$ from $\widetilde{\mathcal{I}_{\mathcal{A}}}$ via Poisson sampling strategy, which is defined by

$$\begin{cases} \widehat{\mathcal{I}^u_{\mathcal{A}}} := \widetilde{\mathcal{I}_{\mathcal{A}}} \odot \boldsymbol{p}^u, \\ \widecheck{\mathcal{I}^u_{\mathcal{A}}} := \widetilde{\mathcal{I}_{\mathcal{A}}} \odot (\mathbf{1} - \boldsymbol{p}^u), \end{cases} \tag{6}$$

in which an independently Poisson sampled instance $\widehat{\mathcal{I}^u_{\mathcal{A}}}$ of the dual self-attention-based noisy image $\widetilde{\mathcal{I}_{\mathcal{A}}}$ with the sampling probability p is defined by

$$\widehat{\mathcal{I}^u_{\mathcal{A}}}[r, c] = \begin{cases} \widetilde{\mathcal{I}_{\mathcal{A}}}[r, c] & \text{if } p, \\ 0 & \text{if } 1 - p. \end{cases} \tag{7}$$

Such that $r \in [1, R]$, $c \in [1, C]$, and $p \in (0, 1)$. By merging Eq. (7), Eq. (3) can be rewritten as

$$\mathcal{L} = \operatorname*{argmin}_{\phi,\xi} \sum_{u=1}^{U} \left|\left| \mathcal{M}_{\phi,\xi}(\widehat{\mathcal{I}^{u}_{\mathcal{A}}}) - \widetilde{\mathcal{I}^{u}_{\mathcal{A}}} \right|\right|^{2}_{\boldsymbol{p}^{u}}, \tag{8}$$

from which we can see that the above loss function is calculated on those sampled pixels within the learned dual self-attention region $\mathcal{A}$ in each pair of images are randomly selected by $\boldsymbol{p}^u$. Such a loss function provides a fair comparison in an ensemble manner, *i.e.* via accumulations of pixel-wise differences over varying sized region $\mathcal{A}$ in all the U pairs.

3.3. Denoising Scheme

As introduced above, Poisson `dropout` operations $\boldsymbol{p}$ were enforced in IDEA-Net to reduce the prediction variance. Concretely, IDEA-Net yields the denoised image $\widehat{\mathcal{I}_{\mathcal{A}}}$ in an ensemble manner, *i.e.* via the average over multiple predictions $\ddot{\mathcal{I}}^{v}_{\mathcal{A}}$ with model weights associated with independently drawn Poisson sampling probability. With the practical solution proposed in [10], denoising scheme is conducted by

$$\widehat{\mathcal{I}_{\mathcal{A}}} = \frac{1}{V} \sum_{v=1}^{V} \ddot{\mathcal{I}}^{v}_{\mathcal{A}} = \frac{1}{V} \sum_{v=1}^{V} \mathcal{M}_{\{\phi,\xi\}^{v}} \widetilde{\mathcal{I}_{\mathcal{A}}} \odot \boldsymbol{p}^{U+v}. \tag{9}$$

4. Experiments

In this section, we first evaluate the performance of IDEA-Net on two denoising tasks: blind Gaussian denoising (*i.e.* AWGN) and real-world noisy image denoising. Then, we further measure its practicability of removing real-world noise in low-light and noisy scenes for face detection. It is noteworthy that we only present partial results in this section due to space limitation. More experimental results are available in our *supplementary materials*.

4.1. Datasets

Datasets for AWGN noise removal. We employ Set14 [42], (C)BSD11, and (C)BSD68 [13, 27] datasets for this task. In particular, we construct the (C)BSD11 dataset by including 2 images from the BSD68 dataset and 9 images from the Colour BSD68 dataset.

Dataset for real-world noise removal. We adopt 40 pairs of noisy images and ground truth images captured by Canon EOS 5D Mark II camera in the PolyU dataset [35] and each of which is with the resolution of 512×512 pixels.

Dataset for face detection in low-light and noisy scenes. In this down-stream task of image denoising, without given ground truth (clear images), we use all the 100 sample testing images from DARK FACE dataset [40] and each of which is with resolution of 1080×720 pixels, and contains 1 to 34 faces within varying sizes of bounding boxes ranging from 1×2 to 335×296 pixels. It is a challenging face detection dataset as it contains a high degree of variability in scale, pose, occlusion, appearance and illumination.

4.2. Evaluation Metric

Metric for image denoising. Peak Signal-to-Noise Ratio (PSNR), as one of the most common full-reference quality metrics, is adopted for all the comparisons over the learned dual self-attention region $\mathcal{A}$ with respect to intensity differences. This is measured by

$$\text{PSNR}\left(\mathcal{I}_{\mathcal{A}}, \widehat{\mathcal{I}_{\mathcal{A}}}\right) = 10 \cdot \log_{10}\left(\frac{255^2}{||\mathcal{I}_{\mathcal{A}} - \widehat{\mathcal{I}_{\mathcal{A}}}||^2_2}\right). \tag{10}$$

Metric for face detection in low-light and noisy scenes. Since human face is the only class in this task, thereby Average Precision (AP) is adopted as performance metric in which Intersection over Union (IoU) is fixed to 0.5. The detection precision is yielded by the official evaluation tool[1].

4.3. Implementation Details

We train the IDEA-Net with TensorFlow 1.14.0 and CUDA 10.0 on a NVIDIA Tesla V100 GPU. As such, our implementation takes ∼10 minutes to process an image with the resolution of 256×256 pixels. Throughout all the experiments, the hyper-parameter of LReLU is respectively valued as 0.2 and 0.1 in $\mathcal{M}_{\phi}$ and $\mathcal{M}_{\xi}$.

In $\mathcal{M}_{\phi}$, the architecture of `A1B` is: `c7s1-32-IN-R, c3s2-64-IN-R, r-64, up-2, c3s1-64-IN-R, up-2, c3s1-32-IN-R, c7s1-1-S`. For each of the five convolution layer, `c` represents `convolution`, `s` denotes stride, and `IN` indicates the Instance Normalisation [30]. Besides, `r-64` indicates residual block operates on 64 channels and `up-2` denotes the nearest neighbour up-sampling with a factor of 2. `R` and `S` respectively represents the ReLU and Sigmoid activation function. The discriminator $\mathcal{D}$ is constructed with the following: `c4s2-64-IN-LR, c4s2-128-IN-LR, c4s2-256-IN-LR, c4s1-512-IN-LR, c4s2-1`. In `A2B` (*i.e.* Algorithm 1), binarisation threshold b and area threshold s are set to be 130 and 500, respectively. $\mathcal{M}_{\phi}$ is trained with 10 iterations.

In $\mathcal{M}_{\xi}$, all the `Conv` and `PConv` layers in the five-block encoder and four-block decoder are with 3×3 kernels, stride of 1, as well as zero-padding of length 2. All the `Conv` and `PConv` layers are activated by LReLU activation function except the last `Conv` layer in which Sigmoid activation function is applied. The `dropout` rate and probability p of the Poisson sampling process are valued as 0.3 and 0.4, respectively. For training $\mathcal{M}_{\xi}$, Adam optimiser is adopted

[1] `https://github.com/Ir1d/DARKFACE_eval_tools`

Dataset	σ	Single-image learning/non-learning						Dataset-based learning		
		KSVD	(C)BM3D	NCSR	TWSC	S2S	Ours	(C)DnCNN	IRCNN	FDnCNN
Set14	25	20.96	28.31	21.99	23.18	30.13	**30.36**	30.03	**30.82**	30.22
	50	16.45	25.02	16.51	19.42	27.72	**27.92**	27.51	**27.80**	27.61
(C)BSD11	25	20.76	29.42	20.11	22.32	29.52	**29.65**	28.58	**30.35**	30.15
	50	15.89	25.57	15.42	18.14	26.56	**26.86**	25.95	**27.03**	26.59
(C)BSD68	25	26.25	28.71	20.32	23.44	29.78	**29.92**	**30.33**	30.26	30.31
	50	23.15	25.46	20.16	19.56	26.97	**27.25**	26.32	27.18	**27.20**

Table 1. Quantitative evaluation of various methods of removing AWGN on the Set14 and BSD68 datasets with different noisy (σ) levels. The metrics are averaged over PSNR (in dB) within our attentional region. The best results in each category of methods under each image-wise noisy level (σ) are marked in bold.

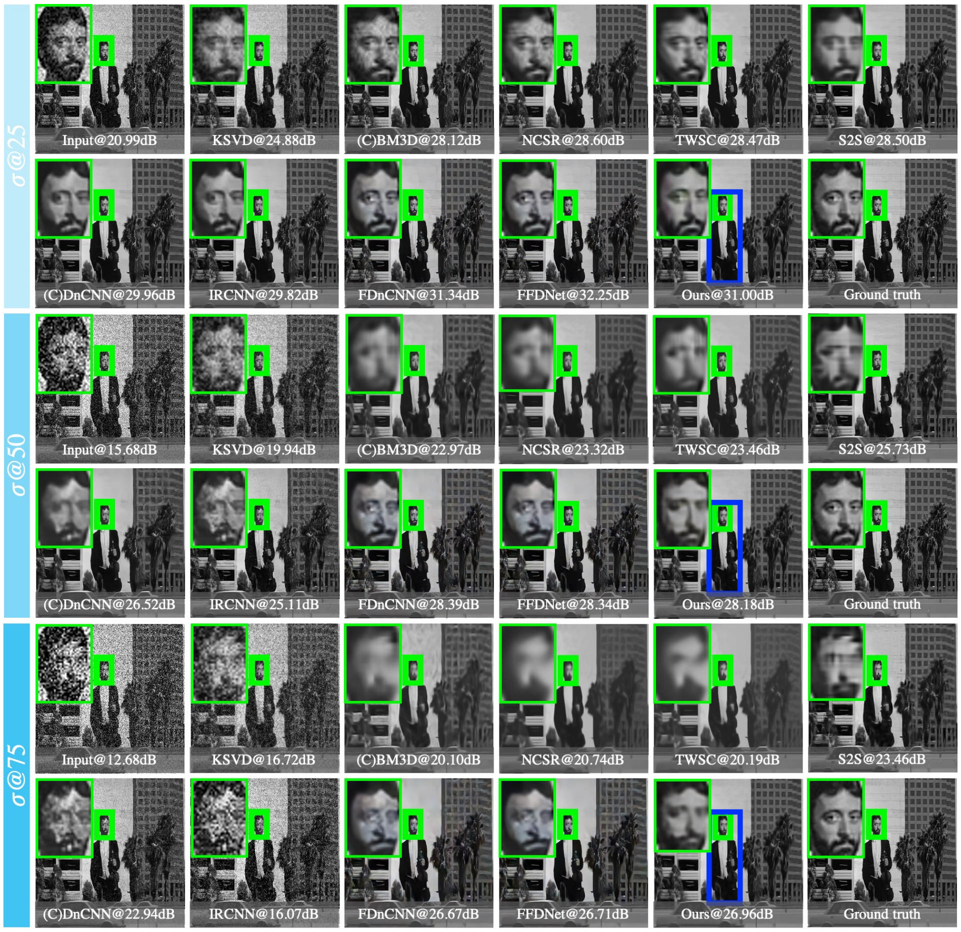

Figure 3. Comparisons of denoising results with respect to PSNR in the case of AWGN with σ valued as 25, 50, and 75. □ denotes the selected image region for comparison and □ indicates the attention $\mathcal{A}$ drawn by IDEA-Net. Best viewed in colour and zoomed mode.

with learning rate initialised as 10^{-5} with 3×10^4 steps. For testing, the number of `dropout` operations is valued as 30.

4.4. Removing AWGN Image Noise

In this task, the noisy level σ is valued as 25 and 50 for each of the three publicly available datasets. The comparative experimental results is quantitatively summarised in Table 1 and one particular example demonstratively visualised in Figure 3.

In comparison with single-image based learning or non-learning methods, the observations are three-fold: 1) our method significantly outperforms KSVD, which reveals the benefits of deep learning compared against dictionary learning; 2) the proposed method is slightly better than existing single-image-based denoisers such as S2S on all the noisy levels; 3) our method still outperforms the leading non-

(C)DnCNN	IRCNN	FDnCNN	MSRResNet	FFDNet	S2S	Ours
36.45	36.29	36.19	36.00	36.22	**36.73**	36.69

Table 2. Quantitative evaluation of various methods of removing real-world noise on the PolyU datasets. The metrics are averaged over PSNR (in dB) within our attentional region. As a reference for comparison, PSNR of the noisy images is 35.47.

learning method (C)BM3D over all the two noisy levels.

In comparison with dataset-based learning methods, this is mainly attributed to the fact that (C)DnCNN has been proved outperformed those methods in [26]. In this experiment, (C)DnCNN, IRCNN and FDnCNN are all pre-trained on BSD300 [21] and (C)BSD300 [13] datasets using noisy levels $\sigma = 25, 50$. Quantitatively, our method outperformed major dataset-based methods and close to the leading one with tiny performance margin.

4.5. Removing Real-World Image Noise

The real-world noisy images are usually resulted from varying camera exposure times [25]. The performance evaluation on real-world noisy image denosing is conducted on the PolyU dataset [35]. All results are summaried in Table 2 and an intuitive comparison is visualised in Figure 4. For the dataset-based methods, the (C)DnCNN, IRCNN, FDnCNN and FFDNet [46] are pre-trained on DND [25] dataset, whereas the super-resolution method MSRResNet [34] is pre-trained on DIV2K [1], Flickr2K [29] and OST [33] datasets. In Table 2, we can observe that our method outperforms all the dataset-based methods and the super-resolution-based one (*i.e.* MSRResNet), and slightly less competitive compared to the state-of-the-art S2S.

4.6. Ablation Study

Time complexity comparisons. As one of the objectives of this paper, we propose IDEA-Net as for the partial sake of reducing the image denoising time. Noting that dataset-based learning methods are incomparable in this study, as they required much different degree of time complexity. In Figure 5, we show that our method requires the shortest time to denoise a total of 11 images in the (C)BSD11 dataset while yields the most competitive PSNR results in both noisy levels.

Convergence rate comparison. To better understand the model stability, we compare IDEA-Net with pioneer single image denoisers DIP and S2S in terms of model convergence rate. Concretely, comparisons are performed in line with optimal PSNR performance obtained with respect to not only its corresponding iteration number but also the time required. Figure 6 confirms that our method required the smallest number of training iterations and the shortest amount of time to reach the optimal PSNR performance.

Downstream task on dark face detection. We use 100 sample testing images from DARK FACE dataset [40]. In particular, given a dark and noise image, the lighting conditions were enhanced by MSRCR [12]. And then, the noises were denoised by selected methods. Finally, face detection was conducted using RetinaNet [18] that pre-trained on WIDER FACE dataset [39]. Note that this dataset does not provide referencing ground truth (*i.e.* clear images). Thereby, dark face detection precision could be treated as a performance metric of image denoising methods. Since no clear images are provided in the DARK FACE dataset, thus blind (C)DnCNN, IRCNN, FFDNet are pre-trained on BSD300 and (C)BSD300 datasets with noisy level within

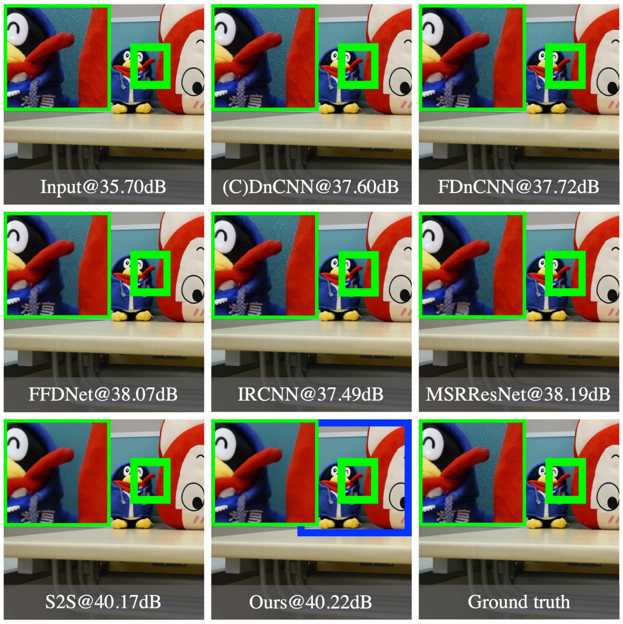

Figure 4. Comparisons of denoising results in terms of PSNR on a real-world noisy image. □ denotes the selected image region for comparison and □ indicates the attention $\mathcal{A}$ drawn by IDEA-Net.

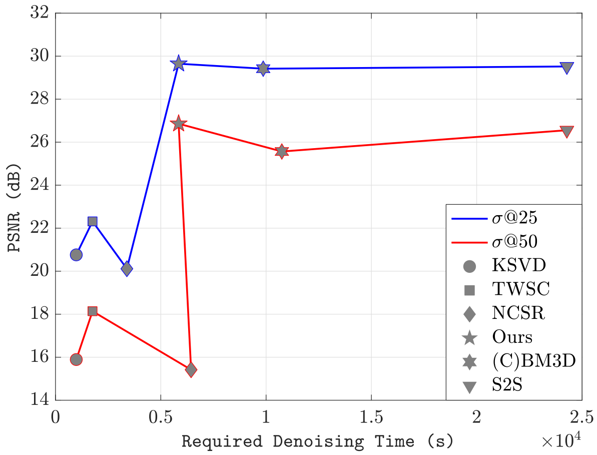

Figure 5. PSNR versus time required for image denoising on the (C)BSD11 dataset under two AWGN levels.

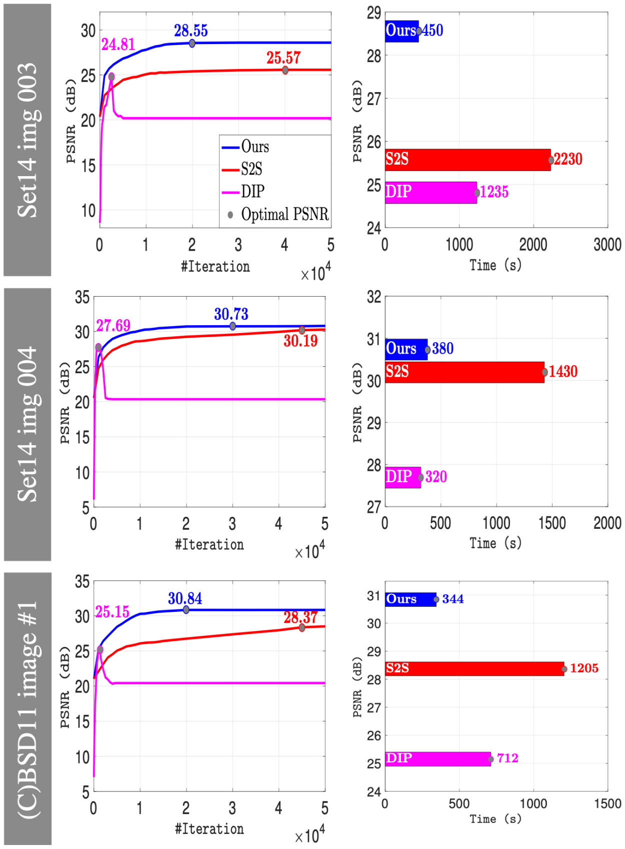

Figure 6. PSNR versus the number of training iterations and time (in seconds) on three images with $\sigma = 25$.

the range of $[0, 55]$. Super-resolution methods MSRResNet, SRMDNF and USRNet are pre-trained on DIV2K, Flickr2K as well as OST datasets. The performance of face detection is summarised in Figure 7. In addition, we also visualise several testing results in Figure 8. In accordance with the summarised results, our method achieves the best face detection precision even though the face detector is pre-trained on the WIDER FACE dataset with normal lighting conditions.

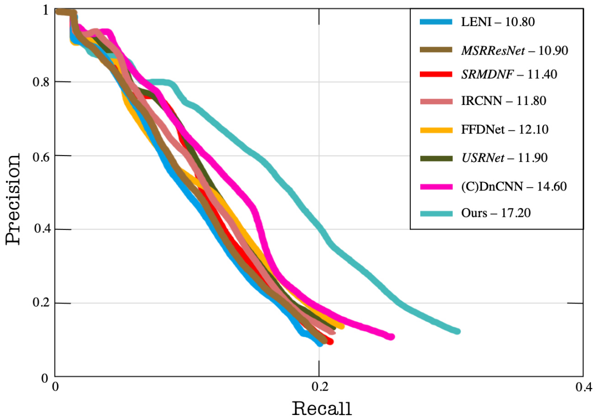

Figure 7. Precision-Recall curves on DARK FACE sample testing subset. Performance is measured by AP (top-right) in %. Super-resolution methods are marked in *italic*. Best viewed in colour.

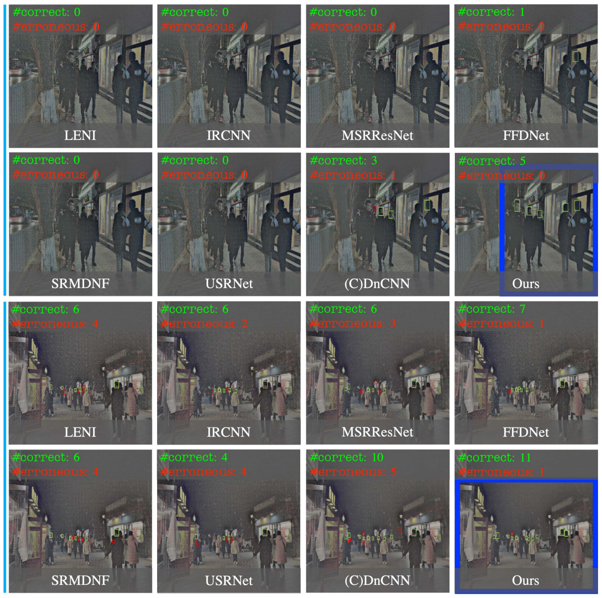

Figure 8. Performance comparisons of dark face detection. □ indicates the attention region $\mathcal{A}$ yielded by IDEA-Net in Figure 2.

5. Conclusion

In this paper, we proposed a self-supervised denoiser IDEA-Net for image denoising. The IDEA-Net requires a noisy image merely for the training process, thus reduces the possible data bias in comparison to those trained using additional datasets. In addition, the learned dual self-attention region in conjunction with Poisson `dropout` operations collectively contribute to the reduction of computational cost and prediction variance. The experimental results show that the proposed IDEA-Net outperforms the non-learning and learning denoisers based on a single image, and is competitive to those trained on datasets. The efficiency and efficacy of IDEA-Net has been further confirmed on the task of face detection in low-light and noisy conditions. Experimental results also inspire further investigations on the spatio-temporal property of dual self-attention-based learning techniques for video denoising.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant No: 61802390), Beijing Natural Science Foundation (Grant No: 4194095), Maoming Science and Technology Program (Grant No: 2020028), Hundred Talents Program of Chinese Academy of Sciences and SunwayAI computation platform.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proc. CVPR Work.*, 2017.

[2] C. Bao, J.-F. Cai, and H. Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. In *Proc. ICCV*, 2013.

[3] J. Batson and L. Royer. Noise2self: Blind denoising by self-supervision. In *Proc. ICML*, 2019.

[4] C. Chen, Z. Xiong, X. Tian, Z. J. Zha, and F. Wu. Real-world image denoising with deep boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(12):3071–3087, 2020.

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

[6] Weisheng Dong, Xin Li, Lei Zhang, and Guangming Shi. Sparsity-based image denoising via dictionary learning and structural clustering. In *Proc. CVPR*, 2011.

[7] W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.*, 22(4):1620–1630, 2012.

[8] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745, 2006.

[9] F. Fan, H. Shan, M. K. Kalra, R. Singh, G. Qian, M. Getzin, Y. Teng, J. Hahn, and G. Wang. Quadratic autoencoder (q-ae) for low-dose ct denoising. *IEEE Trans. Med. Imaging*, 39(6):2035–2050, 2020.

[10] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, 2016.

[11] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. *arXiv preprint arXiv:2101.02824*, 2021.

[12] D. J. Jobson, Z. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.*, 6(7):965–976, 1997.

[13] A. Krull, T.-O. Buchholz, and F. Jug. Noise2void-learning denoising from single noisy images. In *Proc. CVPR*, 2019.

[14] Kanggeun Lee and Won-Ki Jeong. Noise2kernel: Adaptive self-supervised blind denoising using a dilated convolutional kernel architecture. *arXiv preprint arXiv:2012.03623*, 2020.

[15] K. Lee and W.-K. Jeong. Iscl: Interdependent self-cooperative learning for unpaired image denoising. *arXiv preprint arXiv:2102.09858*, 2021.

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

[17] L. Li, X. Yu, Z. Jin, Z. Zhao, X. Zhuang, and Z. Liu. Fdncnn-based image denoising for multi-labfel localization measurement. *Meas.*, 152:107367, 2020.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proc. ICCV*, 2017.

[19] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *Proc. ICCV*, 2009.

[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proc. CVPR*, 2017.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV*, 2001.

[22] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim. Unsupervised attention-guided image to image translation. *arXiv preprint arXiv:1806.02311*, 2018.

[23] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proc. CVPR*, 2020.

[24] C. Niu and G. Wang. Noise2sim–similarity-based self-learning for image denoising. *arXiv preprint arXiv:2011.03384*, 2020.

[25] T. Plotz and S. Roth. Benchmarking denoising algorithms with real photographs. In *Proc. CVPR*, 2017.

[26] Y. Quan, M. Chen, T. Pang, and H. Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proc. CVPR*, 2020.

[27] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proc. CVPR*, 2005.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[29] R. Timofte, R. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proc. CVPR Work.*, 2017.

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proc. CVPR*, 2017.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proc. CVPR*, 2018.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[33] X. Wang, K. Yu, C. Dong, and Chen C. Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proc. CVPR*, 2018.

[34] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proc. ECCV Work.*, 2018.

[35] 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.

[36] J. Xu, L. Zhang, and D. Zhang. A trilateral weighted sparse coding scheme for real-world image denoising. In *Proc. ECCV*, 2018.

[37] B. Yaman, S. A. H. Hosseini, and M. Akçakaya. Noise2inpaint: Learning referenceless denoising by inpainting unrolling. *arXiv preprint arXiv:2006.09450*, 2020.

[38] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Trans. Med. Imaging*, 37(6):1348–1357, 2018.

[39] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proc. CVPR*, 2016.

[40] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Trans. Image Process.*, 29:5737–5752, 2020.

[41] S. Yucer, S. Akcay, N. Al-Moubayed, and T. P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proc. CVPR Work.*, 2020.

[42] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Proc. ICCS*, 2010.

[43] H. Zhang, Y. Li, H. Chen, and C. Shen. Memory-efficient hierarchical neural architecture search for image denoising. In *Proc. CVPR*, 2020.

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

[45] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In *Proc. CVPR*, 2017.

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

[47] H. Zhu and M. K. Ng. Structured dictionary learning for image denoising under mixed gaussian and impulse noise. *IEEE Trans. Image Process.*, 29:6680–6693, 2020.

[48] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.