# Self2Align: a self-supervised denoising framework for single-scene images

Jian Wang

Jian Wang, "Self2Align: a self-supervised denoising framework for single-scene images," Proc. SPIE 12475, Second International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2022), 1247509 (22 November 2022); doi: 10.1117/12.2659707

Event: Second International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2022), 2022, Hulun Buir, China

# Self2Align: A Self-Supervised Denoising Framework for Single-Scene Images

Jian Wang*[a]
[a]Heilongjiang University of Science and Technology, Harbin, China
* Corresponding author: 13782869501@163.com

## ABSTRACT

In recent years, driven by the development of deep neural networks, image denoising technology has made great progress. One of the most representative technologies is the deep neural network denoiser based on supervised learning. The denoiser uses noise-clean image pairs as the input and output of a deep neural network, and trains the deep neural network to achieve the denoising goal. However, collecting many high-quality noise-clean image pairs is extremely challenging, which is mainly because (1) it is difficult to collect true and clean images; and (2) changes in motion and lighting make it impossible to align collected image pairs, which limits the widespread application of supervised learning denoising techniques. To solve the above problems, this paper proposes a simple and effective method Self2Align to train a deep neural network denoiser. First, we proposed an efficient deep network model for inter-image alignment. For the collection of original images, we only use noisy images and collected multiple images for different scenes. The trained alignment network was then used to align the original image pairs automatically. Second, the aligned image pairs generated in the first stage were used as training image pairs for the training of the denoising network. In addition, we introduced a new training strategy so that the network can obtain better performance. The proposed Self2Align architecture eliminates the reliance on noise-clean image pairs and reduces the acquisition difficulty of training image pairs in terms of self-supervised training of the network. We explained the feasibility of our proposed method through theoretical analysis and obtained competitive results through experimental verification.

**Keywords:** Self-Supervised learning, image denoising, U-Net, image align

## 1. INTRODUCTION

An image denoising task involves removing the extra noise from an image that contains noise signals. As the underlying task in the field of computer vision, it is also the core of many image-restoration tasks. The degradation model for the denoising problem is often expressed as:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{x}$ represents the clean image, $\boldsymbol{n}$ represents the additional noise signal, and $\boldsymbol{y}$ is the degraded image affected by noise. Image denoising is a pre-order task of computer vision, and its results profoundly affect subsequent image analysis and processing tasks, such as image recognition and classification. Image denoising methods include traditional image denoising methods and denoisers based on a convolutional neural network (CNN). The traditional denoising methods such as BM3D [1], NLM [2], and other denoising methods based on image local and non-local information are non-learning methods. On the other hand, with the development of deep learning in recent years, CNNs have provided strong support for image processing. CNNs have powerful image feature extraction capabilities and can combine global and local information from images. Researchers have proposed some CNN-based image denoisers including DnCNNs [3], SGNs [4], and U-Net [5], all of which show better performance than traditional denoisers. However, the training of the above CNN-based denoisers usually relies too much on many high-quality noise-clean image pairs, and collecting such datasets is always challenging and costly.

In response to the above problems, a series of self-supervised and unsupervised image denoising methods have been proposed in recent years. This class of methods abandons the reliance on clean images and only focuses on noisy images. The commonly used strategies include collecting multiple noisy images independently for each scene to form a training image pair, which is then used to train the network Noise2Noise [6]. Alternatively, a blind spot network carefully designed will be used to realize the self-supervised learning of the network only relying on a single noisy image. For the self-

Second International Conference on Advanced Algorithms and Signal Image Processing
(AASIP 2022), edited by Kannimuthu Subramaniyam, Proc. of SPIE Vol. 12475,
1247509 · © 2022 SPIE · 0277-786X · doi: 10.1117/12.2659707

supervised learning of Self2Self [7], Noise2Void [8], and the deep network denoiser that does not need training samples for restoration and may restore only from a single image, their theory is based on the depth image prior to DIP [9]. However, in the process of real-world application, it is still difficult to acquire multiple noisy images for the same scene, especially for the acquisition of sports scenes and medical images. In addition, networks based on blind-spot strategies are usually less accurate, computationally more expensive, and require additional post-processing. Finally, denoisers that rely on deep image prior theory are difficult to control and costly when applied to batch image denoising.

In this article, we propose Self2Align, a new self-supervised denoising framework, for the difficult problem of noisy image collection. The framework is designed based on a self-supervised method, which divides the image denoising task into two parts: the generation of training image pairs and the removal of image noise. First, for the collection of noisy images, it is mainly affected by the moving scenes (including moving objects in the scene or the camera movement when taking pictures), resulting in the collected independent noisy images corresponding to different clean ground-truth images. By constructing an alignment network sub-module, we input the collected target noisy image together with the degraded noisy image into the trained alignment network and then output the aligned image which forms a training sample pair with the target image. Second, inspired by Noise2Noise [6], we adopted the above-generated noisy image pairs to train the denoising network. Furthermore, we proposed a new training strategy to resolve the error between the aligned image and the target image, aiming to further optimize the performance of the denoising network. Our proposed framework is independent of any specific denoising network, and denoising networks built with any model can be applied to this framework. In addition, the proposed framework adopts a self-supervised training scheme and does not rely on any clean images (ground-truth). After detailed analysis and experiments, the results demonstrate the effectiveness and advantages of our proposed Self2Align algorithm.

The main contributions of this paper are as follows:

1. To resolve the difficulty of collecting training image pairs for moving scenes, we proposed a new image alignment network, which can achieve fast alignment between collected images without additional settings.
2. From the perspective of theoretical analysis, we proposed an additional training strategy based on the quality difference between the aligned image and the original image, which provides good support for improving the performance of the model.
3. We proposed a novel image denoising framework that does not require clean images as references and only employs noisy image pairs for training. In addition, the method does not rely on any prior information for noise modeling, and it can be used widely in theory.

## 2. RELATED WORK

### 2.1 Image denoising based on supervised learning

In recent years, with the development of deep learning, some denoisers based on deep neural networks have been widely proposed thanks to the excellent expressiveness of deep neural networks. Some of these supervised learning-based denoising methods are particularly outstanding in performance, which generally uses noisy/clean image pairs for network training and optimization. One of the more representative ones is the DnCNN [3] model proposed by Zhang et al., which pioneered the combination of convolutional neural networks and staggered learning modules, and carried out learning under the supervision of noisy/clean image pairs. Compared with the traditional image denoising methods, DnCNN shows greater advantages. Therefore, it is often used as a baseline model for reference. After this, some more prominent denoising networks have been proposed, such as [4, 10, 11, 12]. However, these networks require many noisy/clean image pairs as training data, which is always challenging and expensive in real-world image collection, thus limiting the application of these methods.

### 2.2 Image denoising based on self-supervised learning

Image denoising methods can be divided into traditional denoising methods and deep network-based denoising methods. Traditional denoising methods mainly rely on the pre-designed image prior information or are based on the non-local image prior principles, etc. These methods include BM3D [1] and NLM [2] which are all non-learning methods.

To get rid of the need for clean images in supervised learning, some denoisers that do not rely on clean images have been proposed. Lehtinen et al. proposed Noise2Noise [6], which proved through theoretical analysis that using only noisy image pairs for training will also bring about results comparable to supervised learning. However, the collection of noisy image pairs is still difficult. Noise2Void [8] and Noise2Self [16] which were proposed later adopted a blind spot strategy to avoid identity mapping, further reducing the dependence on noisy images. Self2Self [7] proposed by Yuhui Quan et al. is based on the blind spot idea, which indirectly generates multiple sample copies of noisy images through a delicately designed blind spot network, combining post-processing to improve the final denoising effect of the network. Although the model itself eliminates the dependence on the dataset, the blind spot network requires a large amount of computation and has low accuracy. Dmitry Ulyanov et al. tried to eliminate the explicit noise modeling prior, and instead focused on the implicit prior information for image feature extraction inside CNNs. Their proposed DIP [9] achieved excellent results through a targeted training strategy design. However, the training process of the model itself is difficult to control, and the network needs to be trained separately for each picture, which is inefficient.

# 3. MOTIVATION

In this section, we analyze the theoretical part of the denoising framework proposed in this article. This theoretical framework serves as an important support for Section 4. The main contents of this section are as follows. In Section 3.1, we again focus on the theory proposed in Noise2Noise [6]. This theory holds that using multiple noisy images from the same scene to train the denoising model will bring about the effect of supervised learning. In Section 3.2, based on the above conclusions, to solve the difficult problems in the image acquisition process, we try to model multiple noisy images in the same scene, and mathematically analyze the source of errors during model optimization, as inspired by Neighbor2Neighbor [13]. In Section 3.3, we propose an image alignment method with the help of the image transfer matrix to resolve the error generated during model training, which provides theoretical support for reducing the model training error.

## 3.1 Denoising using only noise/noisy image pairs

The progress of self-supervised denoising algorithms eliminates the dependence on clean images, the most representative of which is Noise2Noise [6] proposed by Lehtinen et al. This method requires that pairs of noisy images are collected for the same scene, and the clean images corresponding to these noisy images are the same. Its objective function is expressed in Equation (2).

$$\arg\min_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{z}} \| f_{\theta}(\mathbf{y}) - \mathbf{z} \|_2^2 \tag{2}$$

where $\mathbf{y}$ and $\mathbf{z}$ represent independent noisy images corresponding to the same clean image (ground-truth) $\boldsymbol{x}$, meeting the relation $\mathbb{E}\{\boldsymbol{y}|\boldsymbol{z}\} = \boldsymbol{x}$; $f_{\theta}(*)$ is the denoising network defined by the parameter $\boldsymbol{\theta}$. By minimizing this loss function, the denoising network can reach an effect similar to supervised learning.

## 3.2 Errors of pairwise noisy image training

According to Section 3.1, we can observe that Noise2Noise only needs multiple noisy images to train the denoising network based on meeting the need for the same clean image (ground-truth). However, in practical applications, it is still difficult and expensive to acquire multiple noisy images for the same scene due to factors such as motion or illumination changes. The main problem is that when the camera is used to take pictures of the same scene to obtain images, the contents of the obtained image pairs are not aligned due to the movement of objects in the scene or the change of illumination, in other words, the pixel values of the corresponding positions are different. An imaging example is shown in Figure 1.

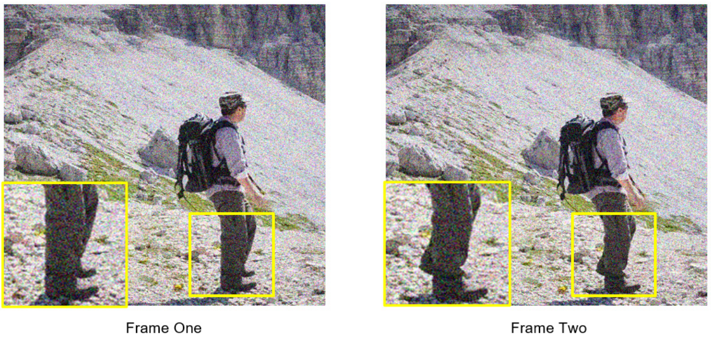


Figure 1. The two frames of images obtained by taking two shots of the same scene under dynamic imaging condition, we can find the difference of image contents between the two frames before and after.

For the inter-image error caused by the above imaging problems, we modeled the error $\varepsilon$ based on the representation in 3.1, and we can get:

$$\varepsilon \coloneqq \mathbb{E}_{\mathbf{z}|\mathbf{x}}(\mathbf{z}) - \mathbb{E}_{\mathbf{y}|\mathbf{x}}(\mathbf{y}) \neq 0 \tag{3}$$

where $\mathbb{E}_{\mathbf{y}|\mathbf{x}}(\mathbf{y}) = \mathrm{x}$ $\mathbb{E}_{\mathbf{z}|\mathbf{x}}(\mathbf{z}) = \mathrm{x} + \varepsilon$. On this basis, we analyze the effect of this error on the training loss. Inspired by the Neighbor2Neighbor [13] proposed by Tao Huang et al., we propose a theorem similar to Section 3.2 in [13]:

$$\mathbb{E}_{\mathbf{x,y,z}}\|f_\theta(\mathbf{y})-\mathbf{z}\|_2^2 = \mathbb{E}_{\mathbf{x,y}}\|f_\theta(\mathbf{y})-\mathbf{x}\|_2^2 - \mathbb{E}_{\mathbf{x,z}} \parallel \mathbf{x}-\mathbf{z} \parallel_2^2 - 2\varepsilon\mathbb{E}_{\mathbf{y,z}}(f_\theta(\mathbf{y})-\mathbf{z}) \tag{4}$$

Through this theorem, we can analyze and get: when $\varepsilon \neq 0$, as $\mathbb{E}_{\mathbf{x,z}} \parallel \mathbf{x}-\mathbf{z} \parallel_2^2 \neq 0$ *and* $\mathbb{E}_{\mathbf{y,z}}(f_\theta(\mathbf{y})-\mathbf{z}) \neq 0$, the loss function optimized based on Noise2Noise $\mathbb{E}_{\mathbf{x,y,z}}\|f_\theta(\mathbf{y})-\mathbf{z}\|_2^2$ cannot reach a comparable effect as the optimized loss function with supervised learning. However, if we try to decrease $\varepsilon$, i.e., $\varepsilon \to 0$, we can know $2\varepsilon\mathbb{E}_{\mathbf{y,z}}(f_\theta(\mathbf{y})-\mathbf{z}) \to 0$. Considering that in $\mathbb{E}_{\mathbf{x,z}} \parallel \mathbf{x}-\mathbf{z} \parallel_2^2$, the variance of $\mathbf{z}$ is a constant, minimizing the objective function $\mathbb{E}_{\mathbf{x,y,z}}\|f_\theta(\mathbf{y})-\mathbf{z}\|_2^2$ of Noise2Noise at this time can achieve the purpose of optimizing the objective function of supervised learning $\mathbb{E}_{\mathbf{x,y}}\|f_\theta(\mathbf{y})-\mathbf{x}\|_2^2$ at the same time. So far, we have analyzed the influence of the error between noisy images on the optimization of the Noise2Noise objective function in detail, and now reducing the value of $\varepsilon$ will be our next goal.

### 3.3 Transfer matrix for noisy image alignment

First, for the same scene, under dynamic imaging conditions, we take pictures in different time frames to obtain multiple noisy images. We model it in the order of time frames as $\mathbf{Y}\{y_0, y_1, y_2, \dots, y_n\}$, where $y_n$ represents the *n-th* noisy image. Assuming $y_0$ is the target image, then $y_0 = x_0 + n_0$. Secondly, considering that the moving objects are imaged non-aligned at different time frames under dynamic imaging conditions, we set $M_n$ as the transfer matrix to measure the noisy image $y_0$ of the target frame to the noisy image $y_n$ of the remaining frames, then we have:

$$y_0 \circ M_i = y_i = x_i + n_i = x_0 \circ M_i + n_0 \circ M_i \quad (i = 1,2,\dots,n) \tag{5}$$

Based on the specific motion model and noise statistical model, the noisy image of each period $\{y_i\}_{i=1}^N$ can be generated by the target frame noisy image $y_0$ through the transfer matrix. Therefore, with the help of the transfer matrix, we can further obtain the relationship between the target frame image and other frame images when they are aligned, and we have:

$$y_0 = y_i \circ M^{-1} = (x_i + n_i) \circ M^{-1} = x_i \circ M^{-1} + n_i \circ M^{-1} \quad (i = 1, 2, \dots, n) \tag{6}$$

The above formula shows that by solving the inverse matrix of the transfer matrix, we can realize the transformation from the image of any frame to the image of the target frame, where $y_0, y_1 \circ M^{-1}, y_2 \circ M^{-1}, \dots, y_n \circ M^{-1}$ is a set of images formed by transforming each frame of images, and the images in this set share the same $x_0$ (ground-truth) and noise distribution. The process of realizing the alignment between images by transfer matrix $M^{-1}$ is shown in Figure 2.

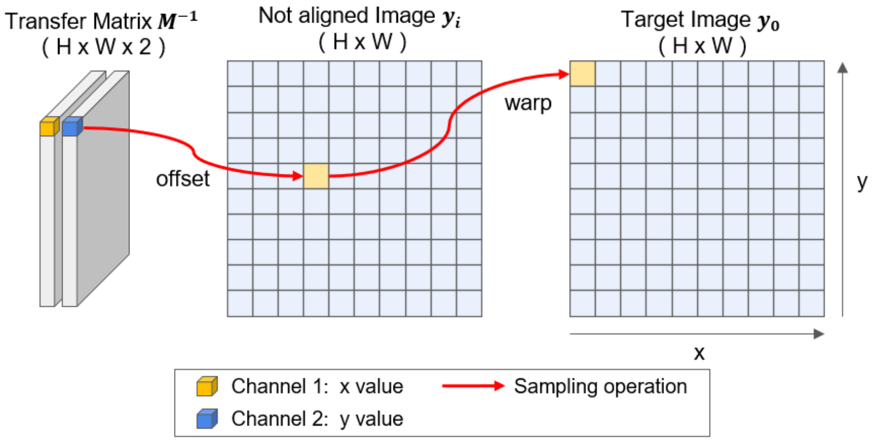


Figure 2. The transfer matrix $M^{-1}$ contains the sampling information of pixels at corresponding positions in the target image $y_0$, and the unaligned image $y_i$ is sampled through the transfer matrix $M^{-1}$ to realize the alignment between images.

Finally, based on the above conclusions, we propose the IAN model to estimate the transfer matrix between images in different frames. Let $F_A(*)$ be the transfer matrix estimation function, we have ${M_i}^{-1} = F_A(y_i, y_0)$, where the image alignment operation is done by the grid sampler, and the aligned image is denoted as $y' = y_i \circ {M_i}^{-1}$. We learn the mapping of the transfer matrix by $F_A(*)$ by constructing the IAN network, using the L2 norm for optimization during network training.

## 4. PROPOSED METHOD

In this section, we propose a new self-supervised denoising framework Self2Align, which consists of two parts. In Section 4.1, we first propose IAN, a deep network architecture for alignment between noisy images, which reduces the inter-image error caused by dynamic imaging. In Section 4.2, while we train the denoising network with paired images generated by the alignment network, we design an additional training strategy for the error $\varepsilon$, which further reduces the loss of the aligned image for ground truth.

### 4.1 Noisy Image Alignment Network IAN

Aiming at resolving the inter-image error generated under dynamic imaging conditions, we propose a deep alignment network IAN, the schematic diagram of the network architecture is shown in Figure 3.

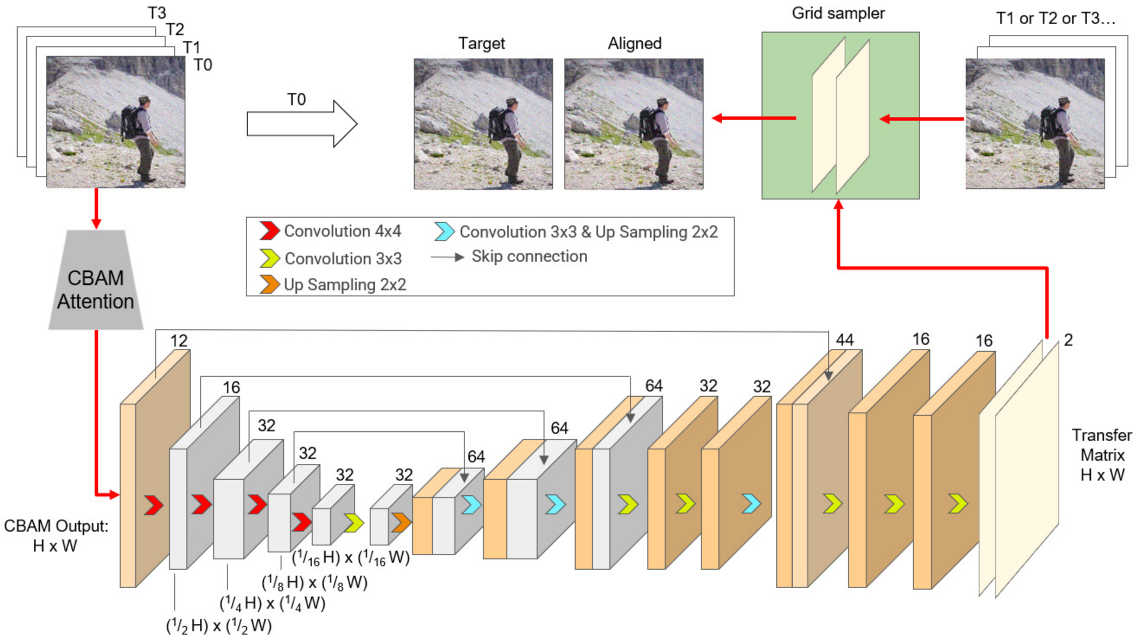


Figure 3. The Architecture of Noisy Image Alignment Network IAN

Overall, the architecture is divided into two parts: encoding/decoding neural network and image sampling. In the encoding/decoding part, a sequence of time frames (T0, T1, ..., Tn) is provided as the network input (the size of each frame image is H × W × 3), and it is assumed that T0 is selected as the target frame. First, the input image is preprocessed by the CBAM Attention [14] module, and the output H × W × 12 feature cube as the first layer input of the encoder is mapped to the H × W ×16 feature cube, and then block processed by the following 3 encoders. Each encoder module consists of a convolutional layer of size 4 × 4, stride 2, and LRELU as the activation function, and the number of channels output by each layer is fixed at 32. The encoder finally outputs a feature cube of H/16 × W/16 × 32. The decoder consists of ten decoder blocks, the first two blocks in turn consist of convolutional layers with size 3 × 3, stride 1, LRELU, upsampling layers with scaling factor 2, and skip connections. The skip connection is used to stack the up-sampled output results with the output results of the corresponding blocks of the encoder by channel. The third decoding block contains both 3 × 3 convolutions, LRELU, up sampling layers, and skip connections. The same structure also appears in the fourth and seventh decoding blocks. The middle two blocks and the last three blocks are both 3 × 3 convolutions, where the number of feature cube channels output by the middle two blocks is fixed at 32. The number of feature cube channels output by the last three blocks gradually decreases. The decoder finally outputs a full-resolution transfer matrix with a channel count of two. In the image sampling part, based on the transfer matrix output by the encoding/decoding network, grid sampling is performed on the images to be aligned [15], and the predicted image aligned with the target frame image is finally output.

### 4.2 Self2Align denoising framework

The proposed Self2Align denoising framework's schematic diagram is shown in Figure 4. The framework mainly includes noisy image alignment network IAN, Bernoulli sampler, and image denoising network. Among them, the Bernoulli sampler accepts the output of IAN and the target frame as input and generates image pairs for denoising network training through a given sampling strategy. For the Bernoulli sampling strategy, we have:

$$\hat{\boldsymbol{y}}[i] = \begin{cases} \boldsymbol{y}[i], & p \\ 0, & 1\text{-}p \end{cases} \quad p: sampling\ probability$$

$Condition\ by$:

$$p = \begin{cases} 1, & 0.5 < \sigma_{SSIM} \le 1 \\ \sigma_{SSIM}, & 0 \le \sigma_{SSIM} \le 0.5 \end{cases} \quad with \quad F_{SSIM}\left(I_{predict}, I_{target}\right) \xrightarrow{yields} \sigma_{SSIM} \tag{7}$$

The above equations describe our proposed sampling strategy, where $\boldsymbol{y}$ is the image before sampling, $\hat{\boldsymbol{y}}$ is the image after sampling, $F_{SSIM}$ is used to calculate the structural similarity coefficient $\sigma_{SSIM}$ between the output image $I_{predict}$ of the alignment network IAN and the target frame $I_{target}$.

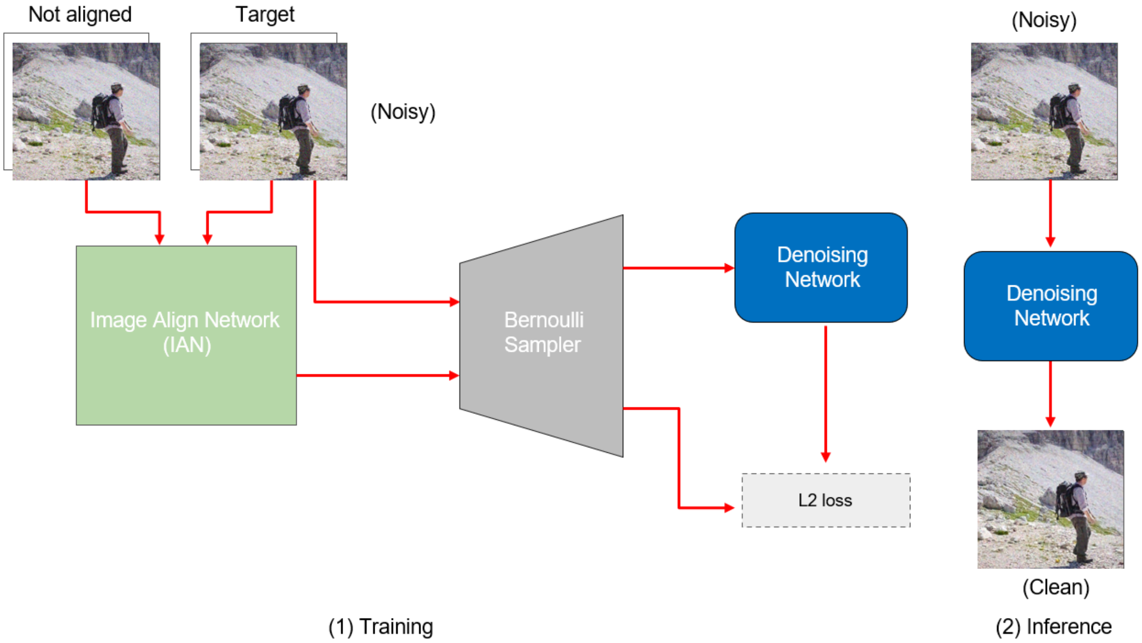


Figure 4. Self2Align denoising framework

We perform this sampling strategy on the aligned image $I_{predict}$ output by IAN and get image pairs $I_{pair} = \left(\hat{I}_{predict}, I_{target}\right)$, which are used for training the denoising network. In the above sampling strategy, we adopt the structural similarity coefficient to measure the quality of the aligned images. It should be noted that there are the following training strategies in the process of denoising network training: when $p = 1$, error $\varepsilon$ of the image to $I_{pair}$ is small enough, and the training process at this time is similar to Noise2Noise [6]. For some extreme imaging conditions, for instance, when $p = \sigma_{SSIM}$, error $\varepsilon$ of the image to $I_{pair}$ is amplified, and the training process at this time is similar to DIP [9]. For the denoising network, we use U-Net, where the blind spot sampling rate of the input layer is set to 0.5 by default, and the loss function is L2.

# 5. EXPERIMENT

In this section, we describe the entire experiment in detail, including the implementation of our denoising method and a comparative analysis with other algorithms. In Section 5.1, we first introduce the training details of Self2Align, including hardware and the adjustment of training parameters, and the dataset used for training. In Section 5.2, we will compare the performance capabilities of our denoising algorithm with other algorithms, including qualitative and quantitative visual comparisons. Finally, in Section 5.3 we will conduct ablation experiments with our proposed training strategy to confirm the effectiveness of our proposed method.

## 5.1 Implementation Details

**Training details:** We trained the alignment network IAN in sRGB space. The training batch size was 90. The number of iterations for training was 120, and the ADAM optimizer was used for parameter optimization. In terms of the learning rate, we set the initial learning rate of the network to 0.004, and the learning rate decayed to 0.8 times the original value every time the network was trained for 10 iterations. For the denoising network, we used the U-Net structure, set the blind spot network in the first layer, and set the blind spot sampling rate to 0.5. The denoising network training batch size was 200, and the ADAM optimizer was also used for 100 iterations of training. The initial learning rate was set to 0.001, and the learning rate was halved every 10 iterations. Among them, all experiments were performed on a server equipped with NVIDIA RTX6000 GPU, Python 3.6, and PyTorch 1.10.

**Experimental dataset:** During the network training process, for the alignment network IAN, we used the DAVIS [17] dataset for training, which contained 9,000 images of 90 scenes. We fixed its size to 448 × 448 by cropping and synthesized the noisy image by applying Gaussian noise. When training the denoising network, we used Vimeo-90K [18] as the original dataset. 30K scene images were selected from it for cropping and alignment processing, and the processed images were used as the training dataset for the denoising network. In the method testing part, we used Kodak [19], BSD300 [20], and Set14 [21] as test datasets.

## 5.2 Comparison of methods

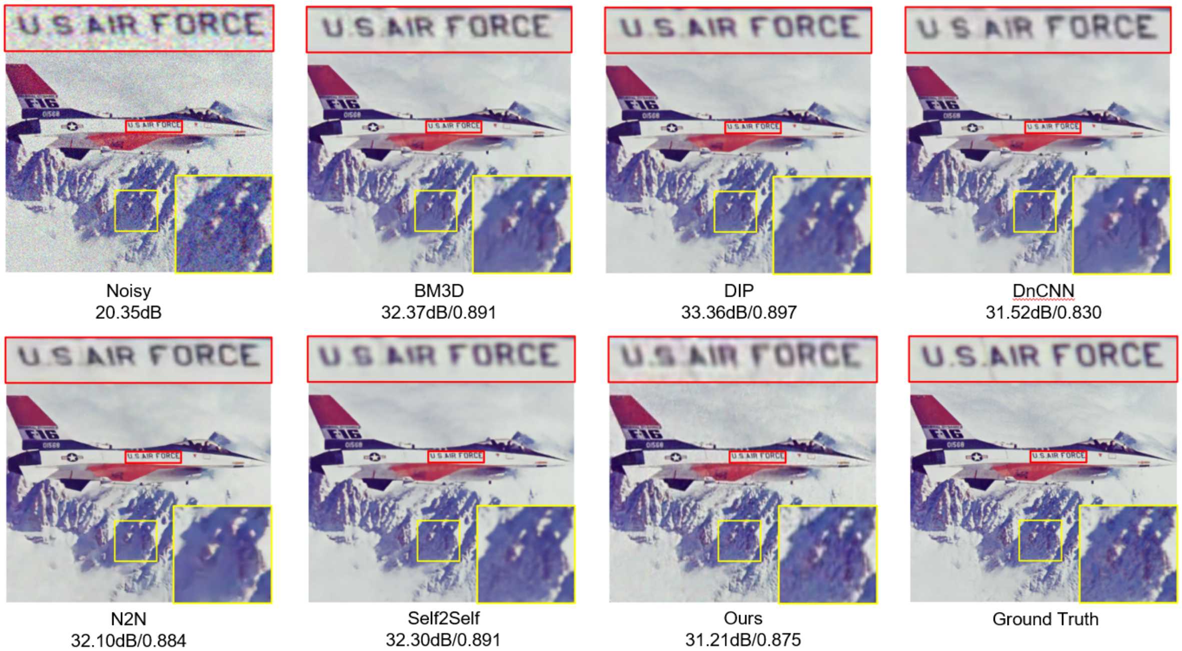
**Experimental details:** To compare the performance capabilities of our proposed Self2Align with other methods, we selected some of the most popular denoising algorithms. Among them, supervised learning DnCNN (N2C) [3] and self-supervised learning Noise2Noise (N2N) [6] were used as baseline methods. There is also another traditional denoising algorithm BM3D [1] and two self-supervised denoising algorithms Deep Image Prior (DIP) [9] and Self2Self [7]. Among them, for DnCNN and Noise2Noise, we used pre-trained weights for experiments, and for the performance of BM3D, DIP, and Self2Self we refer to the results provided by [13].

**Experimental results:** In comparative experiments, we used PSNR and SSIM as quantitative standards to compare the performance of denoising networks. The experimental results are shown in Table 1. It can be seen from the experimental results that our Self2Align is completely better than DIP in performance, and our method also performs better on the BSD300 dataset compared to the supervised learning DnCNN. For Self2Self trained on a single image, our method also achieves competitive results, especially in image structure restoration. One of the possible reasons is that we used the image alignment network IAN to pre-align the training images, thus preserving the prior information of the image structure as much as possible, which can also be verified in the non-learning BM3D and Noise2Noise. In addition, from the sampling strategy in Section 4.2, we know that our denoising network will perform a training process similar to N2N and DIP based on the quality of the aligned images. From Section 3.2, we know that the network optimization based on N2N will indirectly optimize supervised learning weights of the denoiser, and this conclusion is also reflected in the similarity of the experimental results of DnCNN and the proposed method. The relationship of DIP to our method and the impact of sampling strategy on the performance of our method is described in detail in Section 5.3.

The visual performance of different methods is shown in Figure 5. From the figure, we can find that our method achieves competitive results in both quantitative and qualitative aspects. Our method is close to Self2Self in image structure and details and is more similar in quantitative results than N2N. Furthermore, our method has more image details, which may be due to the structural similarity priors introduced in the sampling strategy. Compared with the denoising results of DnCNN, it can be found that our method is very close to or even slightly better than it, which further verifies our above analysis of the network training loss. The good performance of DIP on image quality mainly relies on the long-term training and multiple iterations of the model, coupled with the need to carefully control the nodes where the training stops, which limits the widespread use of this method.

Table 1. Denoising results obtained by adding Gaussian noise $(\sigma = 25)$ on KODAK, BSD300 and SET14 are expressed as average PSNR (dB)/SSIM (1.00E-1). The best results based on deep learning methods are marked in bold letters, and the second-best results are marked with an underline.

| Dataset | BM3D | DIP | Self2Self | Ours | Baseline | |
|---|---|---|---|---|---|---|
| | | | | | DnCNN | N2N |
| **KODAK** | 31.87/0.868 | 27.20/0.720 | **31.28/0.864** | 28.63/0.795 | 30.45/0.909 | 32.41/0.884 |
| **BSD300** | 30.48/0.861 | 26.38/0.708 | **29.86/0.849** | 29.69/0.840 | 29.03/0.895 | 31.04/0.878 |
| **SET14** | 30.88/0.854 | 27.16/0.758 | **30.08/0.839** | 28.53/0.807 | 30.20/0.918 | 31.37/0.868 |



Noisy
20.35dB

BM3D
32.37dB/0.891

DIP
33.36dB/0.897

DnCNN
31.52dB/0.830

N2N
32.10dB/0.884

Self2Self
32.30dB/0.891

Ours
31.21dB/0.875

Ground Truth

Figure 5. At the level of Gaussian noise $(\sigma = 25)$, we compared our method with other competing methods in terms of visual quality, with the quantitative value PSNR (dB)/SSIM below each figure. (* The denoising results of BM3D and DIP are from their project websites [22], [23])

### 5.3 Ablation Study

To investigate the impact of the training strategy proposed above on our method, we will conduct ablation experiments in this section. Since our training strategy is mainly affected by the sampling strategy, we will discuss the two aspects of using different sampling strategies and the reasons that affect the training results.

(1) First, in the input layer of the denoising network, we use the blind spot network by default, which corresponds to Eq. (7) $p \equiv 0.5$, denoted as strategy 1.

(2) Second, we design another sampling strategy as a reference, corresponding to Eq. (7) $p \equiv F_{SSIM}(I_{predict}, I_{target})$, denoted as strategy 2.

Combining the above two situations, we compared the training strategies proposed in this paper experimentally, and the results are shown in Table 2. Through the results in the table, we can find that our proposed training strategy *Strategy Ours* is completely better than the default training strategy 1 and our additionally designed training strategy 2. For strategy 1, the blind spot strategy is adopted. Through DIP [9] and N2V [8], we can know that the network can directly learn the clean content in the image when the noise distribution is independent, to effectively remove the noise in the image. In addition, for the case where the difference between the training noisy image pairs is too large, the input image can be regarded as disordered noise to a certain extent, and the network optimization process can be approximately regarded as DIP. For strategy 2, we introduce the structural similarity prior based on strategy1, which makes the sampling process more flexible. For *Strategy Ours,* when $p = 1$, it was shown that the difference between the training image pairs at this time was small, and the optimization process for network training was similar to N2N [6]; when $p = \sigma_{SSIM}$, it was shown that the difference between training image pairs at this time was quite large, and the optimization process of the network degenerates to DIP. Since our proposed training strategy considers both N2N and DIP optimization processes, the denoising effect of our network is significantly improved compared to the default strategy, and the experimental results further verify its effectiveness.

Table 2. By adding Gaussian noise $(\sigma = 25)$ on KODAK, BSD300, and SET14, the results of training the denoising network using each strategy are expressed as average PSNR (dB)/SSIM (1.00E-1). The best results are marked in bold letters, and the second-best results are marked as underlined.

| **Train Strategy** | | **Data Sets** | |
|---|---|---|---|
| | **KODAK** | **BSD300** | **SET14** |
| **Strategy Ours** | **28.63/0.795** | **29.69/0.840** | **28.53/0.807** |
| **Strategy2** | 25.83/0.713 | 26.41/0.753 | 25.44/0.747 |
| **Strategy1** | 24.21/0.669 | 24.66/0.718 | 23.45/0.705 |

## 6. CONCLUSION

In this paper, we propose a new denoising framework Self2Align, which does not require clean images or noise modeling priors, but only needs aligned pairs of noisy images for self-supervised training of denoising networks, thus effectively addressing the issue of image denoising. We also developed a noisy IAN to help collect the training image pairs under dynamic conditions. To enhance the performance of the network, we designed a new training strategy combining the two optimization processes. The research results showed a certain degree of competitiveness over the comparative methods after detailed theoretical analysis and experimental verification. In the future, we will focus on exploring training strategies and testing the performance of denoising networks under extreme conditions.

## REFERENCES

[1] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," in IEEE Transactions on Image Processing, vol. 16, no. 8, pp. 2080-2095, Aug. 2007, DOI: 10.1109/TIP.2007.901238.

[2] A. Buades, B. Coll and J. -. Morel, "A non-local algorithm for image denoising," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 60-65 vol. 2, DOI: 10.1109/CVPR.2005.38.

[3] Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Trans Image Process. 2017 Jul;26(7):3142-3155. DOI: 10.1109/TIP.2017.2662206. EPUB 2017 Feb 1. PMID: 28166495.

[4] Gu S, Li Y, Gool L V, et al. "Self-Guided Network for Fast Image Denoising" 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019.

[5] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

[6] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. In J. Dy, & A. Krause (Eds.), 35th International Conference on Machine Learning, ICML 2018 (Vol. 7, pp. 4620-4631). (Proceedings of Machine Learning Research; No. 80).

[7] Quan, Y., Chen, M., Pang, T., & Ji, H. (2020). Self2Self With Dropout: Learning Self-Supervised Denoising from Single Image. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1887-1895.

[8] A. Krull, T. Buchholz, and F. Jug, "Noise2Void - Learning Denoising from Single Noisy Images," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2124-2132, Doi: 10.1109/CVPR.2019.00223.

[9] Ulyanov, D., Vedaldi, A., & Lempitsky, V.S. (2018). Deep Image Prior. International Journal of Computer Vision, 128, 1867-1888.

[10] Mao, X., Shen, C., & Yang, Y. (2016). Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. NIPS.

[11] Tai, Y., Yang, J., Liu, X., & Xu, C. (2017). MemNet: A Persistent Memory Network for Image Restoration. 2017 IEEE International Conference on Computer Vision (ICCV), 4549-4557.

[12] Lefkimmiatis, S. (2017). Universal Denoising Networks: A Novel CNN-based Network Architecture for Image Denoising. ArXiv, abs/1711.07807.

[13] Huang, T., Li, S., Jia, X., Lu, H., & Liu, J. (2021). Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14776-14785.

[14] Woo, S., Park, J., Lee, J., & Kweon, I. (2018). CBAM: Convolutional Block Attention Module. ECCV.

[15] Ma, F., Zhu, L., Yang, Y., Zha, S., Kundu, G., Feiszli, M., & Shou, Z. (2020). SF-Net: Single-Frame Supervision for Temporal Action Localization. ArXiv, abs/2003.06845.

[16] Batson, J.D., & Royer, L.A. (2019). Noise2Self: Blind Denoising by Self-Supervision. ArXiv, abs/1901.11365.

[17] Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbelaez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 Davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675.

[18] Video Enhancement with Task-Oriented Flow. source:http://toflow.csail.mit.edu/.

[19] Rich Franzen. Kodak lossless true color image suite. source:http://r0k.us/graphics/kodak.

[20] 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,” Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 2001, pp. 416-423 vol.2, DOI: 10.1109/ICCV.2001.937655.

[21] Zeyde R, Elad M, Protter M. On Single Image Scale-Up Using Sparse-Representations[J]. International Conference on Curves and Surfaces, 2010.

[22] Lebrun, M. (2012). An Analysis and Implementation of the BM3D Image Denoising Method. Image Process. Line, 2, 175-213.

[23] Deep Image Prior. source: https://dmitryulyanov.github.io/deep_image_prior.