Optimal Transport for Unsupervised Denoising Learning

Wei Wang, Fei Wen*, Senior Member, IEEE, Zeyu Yan*, and Peilin Liu, Member, IEEE

Abstract—Recently, much progress has been made in unsupervised denoising learning. However, existing methods more or less rely on some assumptions on the signal and/or degradation model, which limits their practical performance. How to construct an optimal criterion for unsupervised denoising learning without any prior knowledge on the degradation model is still an open question. Toward answering this question, this work proposes a criterion for unsupervised denoising learning based on the optimal transport theory. This criterion has favorable properties, e.g., approximately maximal preservation of the information of the signal, whilst achieving perceptual reconstruction. Furthermore, though a relaxed unconstrained formulation is used in practical implementation, we prove that the relaxed formulation in theory has the same solution as the original constrained formulation. Experiments on synthetic and real-world data, including realistic photographic, microscopy, depth, and raw depth images, demonstrate that the proposed method even compares favorably with supervised methods, e.g., approaching the PSNR of supervised methods while having better perceptual quality. Particularly, for spatially correlated noise and realistic microscopy images, the proposed method not only achieves better perceptual quality but also has higher PSNR than supervised methods. Besides, it shows remarkable superiority in harsh practical conditions with complex noise, e.g., raw depth images. Code is available at https://github.com/wangweiSJTU/OTUR.

Index Terms—Restoration, denoising, unsupervised learning, optimal transport, generative adversarial nets, depth image denoising

1 INTRODUCTION

Image restoration is a fundamental problem in low-level computer vision, which is crucial for many subsequent high-level tasks. The past few years have witnessed great progress in image restoration, which largely benefits from sophisticated deep neural networks and large amounts of paired training data [1], [2], [3]. However, in many real-world applications, it is difficult or expensive to collect noisy-clean image pairs for supervised learning. In this case, synthesized data can be used, but the gap between synthesized and real-world data would fundamentally limit the performance of the learned restoration model on realistic data.

Traditional restoration methods, which do not rely on data driven learning, are typically designed by exploiting certain structures of the signal or properties of noise. Widely used signal structures include sparsity on certain basis, low-rankness, spatiotemporal smoothness, or self-similarity, while popular noise properties include Gaussianity, or spatially independent i.i.d. noise. Particularly, the representative non-local means (NLM) methods, e.g., BM3D [4], [5], have demonstrated excellent effectiveness in denoising i.i.d. Gaussian noise. However, the performance of these methods is limited when the prior assumption on the signal and noise deviates from real-world data. Moreover, the performance of these methods is closely related to hyper-parametric setting, e.g., the degree of sparsity, low-rankness or smoothness. More recently, deep image prior (DIP) [6] has been proposed to use a network as handcrafted prior, which has achieved impressive results but requires optimizing the network for each noisy image.

To address the difficulty in collecting paired noisy-clean training data in some practical applications, data driven learning methods without requiring paired training data have recently attracted much attention. For example, the noise-to-noise (N2N) method [7] learns to restore images by only looking at corrupted noisy samples. To further remove the requirement of paired noisy training data in N2N, the noise-to-void (N2V) and noise-to-self (N2S) methods [8], [9] learn restoration models in a self-supervised manner. Moreover, a number of variants of these methods have been developed, e.g., [10], [11], [12], [13], [14].

Though much progress has been made in restoration learning in the condition without paired training data, existing methods more or less rely on some assumptions of the signal and/or degradation model. For example, the N2N method [7], as well as [10], [12], requires paired noisy training data, whilst the self-supervised methods [8], [9], [13], [14] assume that the noise is spatially independent and/or the noise type is a priori known. Hence, the performance of these methods is limited in some practical applications, especially when the noise or degradation is complex.

Unsupervised restoration learning in the absence of any prior knowledge on the degradation model is still a challenging problem. This work is motivated by the following open question:
How to construct an optimal criterion for unsupervised restoration learning in the absence of any prior knowledge on the degradation model?

Toward answering this question, this work proposes an unsupervised learning criterion for restoration from the optimal transport theory. The proposed criterion not only has good theoretical properties, but also has favorable practical performance even compared with supervised methods in denoising various synthetic and real-world images.

The main contributions of this work are as follows.

First, we propose an unsupervised criterion for restoration learning based on the optimal transport theory, which minimizes the transport cost between the input and reconstruction output under the constraint that the reconstruction output has the same distribution as natural clean samples.

Second, we compare the proposed criterion with the ideal supervised criterion to analyze its properties, which shows that it can well preserve the information of the signal whilst achieving perceptual reconstruction.

Third, the proposed criterion is relaxed in implementation with the use of GAN. In theory we show that, with Wasserstein-1 distance, the relaxed formulation has the same solution as the original criterion.

Finally, we apply the new method to various denoising applications, including synthesized images with different noise conditions, and real-world photographic, microscopy, depth, and raw depth images. The results demonstrate that the new method even compares favorably with supervised methods, with PSNR approaching that of supervised methods while having better perceptual quality. Noteworthy, for spatially correlated noise and realistic microscopy images, it not only achieves better perceptual quality to preserve more details but also has higher PSNR than supervised methods. Particularly, in denoising raw depth images with severe complex noise, the new method shows remarkable performance with significant superiority over existing methods.

Our method not only does not use any hand-crafted model for the signal or any prior knowledge on the noise, but also has sound theoretic foundation from the optimal transport theory. The proposed criterion can serve as an optimal criterion for unsupervised restoration learning in the absence of any prior knowledge on the degradation model.

2 RELATED WORK

We review major categories of restoration methods which do not require paired training data, along with GAN based methods that related to this work.

Hand-crafted model based methods. Most traditional model based methods do not rely on data driven learning but use prior model on the signal and/or noise, e.g., sparsity, low-rankness, spatio-temporal smoothness, or self-similarity of the signal [15], [16]. For example, traditional image filters, such as Gaussian, median, bilateral and other filters, use local averaging to reduce the noise level. The NLM method [4] estimates the central pixel of each patch by weighted average of the central pixels from similar patches. The BM3D method [5] filters out the high-frequency parts of stacked similar patches to achieve more robust denoising. The performance of these methods relies heavily on hyperparameter selection [17]. More recently, the DIP [6] method uses a network as handcrafted prior.

Supervised learning with noisy pairs. Instead of using noisy-clean image pairs, the N2N method [7] uses noisy-noisy pairs for training. In principle, for zero-mean noise, N2N can achieve the same performance as supervised learning using clean targets. Extension of N2N to video denoising has been considered in [10]. When paired noisy images are not available, Noiser2Noise [18] and Noisy-as-clean [19] adopt a strategy that adds synthetic noise to noisy images to construct noiser-noisy pairs for training. Moreover, the GAN2GAN method [20] uses GAN to simulate the noise and then generates training pairs for supervised restoration learning.

Self-supervised methods. The self-supervised methods N2V [8] and N2S [9] only use noisy images for training without requiring noisy pairs. Specifically, such methods use a blind-spot network to predict the masked pixel by neighboring pixels. Besides, additional information such as noise model can be incorporated to further improve the performance [13], [14], [21]. Moreover, the Self2Self method [11] uses Bernoulli sampling to construct paired training data from noisy images, whilst the Neighbor2Neighbor method [22] proposes a random neighbor down-sampler to generate the training pairs. These methods commonly use the assumption of independent noise between pixels and the smoothness of natural images. The independent noise assumption limits the application in complex real-world noise conditions.

Unsupervised learning with Stein’s unbiased risk estimator (SURE). In [23], SURE has been applied to unsupervised training of Gaussian denoiser, which has been shown to perform comparably with methods trained with ground-truth. An extension to the case that paired noisy images are available has been considered in [12]. These SURE-based methods are developed under the assumption that the noise is i.i.d. Gaussian.

Conditional GAN (CGAN) based methods. Naturally, the denoising problem can be treated as a conditional generation problem, i.e., generating a natural clean image conditioned on a noisy input by CGAN [24]. However, in the absence of paired training data, CGAN without supervision can disregard the input and randomly output natural samples far from the input. Accordingly, the RoCGAN [25] uses ground-truth data for supervised training of generation from noisy images. Though AmbientGAN [26] does not rely on paired training data for conditional generation, it assumes that the degradation model is known (pre-defined) and easy to sample. NR-GAN [27] can ease this requirement, but it still requires either the noise distribution type is a priori known, e.g., zero-mean Gaussian, or the noise is transformation invariant.

As introduced above, there exist a large number of methods which do not rely on paired noisy-clean training data. However, existing methods more or less rely on some specific models or assumptions on the signal and/or the degradation model. In comparison, our method does not use any hand-crafted model for the signal or any prior knowledge on the degradation model.
of the deviation between the distributions of restored and
natural samples, e.g., \(d(p_X, p_{\hat{X}})\) [34]. In light of this
understanding, to achieve perceptual reconstruction, a natural
way is to minimize the deviation \(d(p_X, p_{\hat{X}})\) using GAN [34],
[35], [36], [37]. The constraint \(d(p_X, p_{\hat{X}}) \leq 0\) enforces that
the reconstruction has perfect perception quality, i.e., looks like
natural samples from human’s perception.

For ease of implementation, the constrained formulation (2) is typically relaxed to an unconstrained formulation as

\[
\max_f I(f(Y), X) + \lambda d(p_X, p_{\hat{X}}).
\]

(3)

In practical applications, distortion plus adversarial loss
based formulation is typically used for the implementation of (3) as

\[
\min_f L_d(f(Y), X) + \lambda L_{adv}(p_X, p_{\hat{X}}),
\]

(4)

where \(L_d\) is a distortion loss such as MSE, \(\ell_1\)-norm, or
distance between deep features. \(L_{adv}\) is an adversarial loss mea-
suring the deviation between the two distributions \(p_{\hat{X}}\) and \(p_X\)
(e.g., measured by a discriminator). \(\lambda\) is a positive parameter
that balances between the distortion and adversarial losses.

When \(\lambda = 0\), it reduces to the widely used supervised
method without considering perception quality as

\[
\min_f L_d(f(Y), X),
\]

(5)

which is the most popular formulation for supervised resto-
ration learning. Recent studies have shown that minimizing
distortion alone does not necessarily lead to high perceptual
quality, while incorporating perception loss would lead to
increase of the lowest achievable distortion, e.g. MSE [34].

In the formulations (2), (3), (4), and (5), paired training
data \(\{(Y, X)\}\) is required for supervision. In many practical
applications, it is difficult to collect paired training data for
supervised training. For such applications, synthesized
paired data is widely used. However, the gap between syn-
thesized and real-world data would fundamentally limit
the performance on realistic data, especially when the real-
world degradation is too complex to accurately simulate.

To address this problem, numerous methods that do not
require noisy-clean paired training data have been devel-
oped, as detailed in Section 2. Though much progress has
been made recently, the success of most of these methods
relies heavily on certain specific assumptions on the degra-
dation, e.g., i.i.d. noise, spatially independent noise, or the
noise type is a priori known. Such assumptions limit the
practical performance on real-world data with complex
degradation.

### 4 PROPOSED UNSUPERVISED CRITERION FROM
OPTIMAL TRANSPORT

This section first presents the proposed criterion for unsu-
ervised restoration learning. Then, it is relaxed into an
unconstrained formulation for ease of implementation. Par-
ticularly, we show that in theory the relaxed formulation
has the same solution as the original one. Furthermore,
some favorable properties of the proposed formulation are
analyzed in an information-theoretic view.
4.1 An Unsupervised Criterion from Optimal Transport

The optimal transport problem aims to find the most efficient transport map of transforming one distribution of mass to another whilst minimizing the cost, which has wide applications in signal processing, image processing, and machine learning [38], [39]. It can be traced back to the seminal work of Monge [40] in 1781, with significant advancements by Kantorovich [41] in 1942.

Let \( \mathcal{P}(X) \) and \( \mathcal{P}(Y) \) be two sets of probability measures on \( X \) and \( Y \), respectively. Let \( c : Y \times X \rightarrow [0, +\infty) \) be a cost function that \( c(y, x) \) measures the cost of transporting \( y \in Y \) to \( x \in X \). The optimal transport problem seeks the most efficient transport from \( v \sim \mathcal{P}(Y) \) to \( \mu \sim \mathcal{P}(X) \) that satisfies the minimal cost.

**Definition 1.** (Transport map): \( f : Y \rightarrow X \) is a transport map that transports \( v \sim \mathcal{P}(Y) \) to \( \mu \sim \mathcal{P}(X) \), if
\[
\mu(A) = v(f^{-1}(A)),
\]
for all \( \mu \)-measurable sets \( A \).

**Definition 2.** (Monge’s optimal transport problem): For two probability measures \( v \sim \mathcal{P}(Y) \) and \( \mu \sim \mathcal{P}(X) \), find a \( v \)-measurable map \( f : Y \rightarrow X \) by
\[
\inf_{f} \int_{Y} c(f(y), y) dv(y)
\]
subject to \( \mu = f_{\#} v \),

with \( f_{\#} v \) denotes the transport of \( v \) by \( f \). A minimum \( f^* \) to this problem, if there exists one, is called an optimal transport map.

Intuitively, the optimal transport problem finds a transport to turn the mass of \( v \) into \( \mu \) at the minimal cost measured by the cost function \( c \), e.g. \( c(y, f(y)) := \|f(y) - y\|^\beta \)
with \( \beta \geq 1 \).

For the reconstruction problem defined in Section 3, and in the absence of paired training data, we propose an unsupervised learning formulation as
\[
\min_{f} \mathbb{E}_{Y \sim p_Y} (\|Y - f(Y)\|^{\beta})
\]
subject to \( p_\tilde{X} = p_X \),

(7)

where \( \beta \geq 1 \). It is easy to see that this problem is an implementation of the Monge’s optimal transport problem (6).

At first glance, formulation (7) appears counter-intuitive since the objective pushes the reconstruction to the noisy input. However, an in-depth analysis can reveal its favorable properties. Specifically, given the degraded observation \( Y \), it seeks a reconstruction map \( f \) satisfying that:

i) **Perceptual reconstruction**: the constraint \( p_\tilde{X} = p_X \) ensures the reconstruction \( \{\tilde{X}\} \) has the same distribution as the clean natural samples \( \{X\} \), hence has good perception quality, i.e., looks like natural samples.

ii) **Minimum distance transport**: the objective in (7) imposes the fidelity of the reconstruction to the observation \( Y \), by which the formulation has a minimum distance transport property. As will be shown in the next subsection, this property makes the reconstruction map to achieve approximately maximal preservation of the information of \( X \) contained in \( Y \).

These properties make optimal transport a good criterion for unsupervised reconstruction learning in the circumstance that clean natural samples and degraded samples are available but paired noisy-clean data is unavailable. These properties accord well with the goal of perceptual reconstruction as mentioned in Section 3, while without the need of paired training data for supervision.

For ease of implementation, we relax the constrained formulation (7) into an unconstrained one as
\[
\min_{f} \mathbb{E}_{Y \sim p_Y} (\|Y - f(Y)\|^{\beta}) + \lambda d(p_X, p_\tilde{X}),
\]

(8)

where \( d(p_X, p_\tilde{X}) \) measures the deviation between \( p_\tilde{X} \) and \( p_X \), and \( \lambda > 0 \) is a balance parameter. Under mild conditions, this formulation with a penalty parameter \( \lambda \) is equivalent to the formulation with a constraint \( d(p_X, p_\tilde{X}) \leq \mu_\lambda \) for some \( \mu_\lambda > 0 \). As \( \lambda \) increases, the value of \( \mu_\lambda \) decreases. As \( \lambda \rightarrow \infty \), the solution of (8) satisfies that \( d(p_X, p_\tilde{X}) \rightarrow 0 \), i.e. \( p_\tilde{X} \rightarrow p_X \), and the first fidelity term in (8) can be ignored since its weight, which is proportional to \( 1/\lambda \), tends to be 0. In this case formulation (8) can be implemented as GAN, i.e., to generate natural samples by
\[
\min_{f} d(p_X, p_\tilde{X}).
\]

(9)

However, it does not satisfy the minimum distance transport property. For example, the generator can disregard the input \( Y \) and randomly output samples from the distribution \( p_X \) to satisfy \( p_\tilde{X} = p_X \). As a consequence, the proposed method can significantly outperform GAN, as will be shown in experiments.

It is worth noting that, the standard CGAN [24] can avoid this degeneration problem of generating samples far away from \( Y \) by discriminating between data pairs \( (Y, X) \) and \( (Y, \tilde{X}) \), but it requires paired training data for supervision. Similarly, RoCGAN [25] and the distortion plus adversarial loss based method (4) also require paired training data. AmbientGAN does not require paired training data and can be used for conditional generation, but it assumes that the degradation function is known (pre-defined) and easy to sample [26]. This limitation is alleviated in NR-GAN [27], but it requires either the noise distribution type is a priori known (e.g., zero-mean Gaussian) or the noise is rotation, channel shuffle, or color-inverse-invariant. In contrast, though also based on GAN for training, our method avoids the above limitations.

Though relaxed, next we show that the formulation (8) has good theoretical property, e.g., it has the same solution as the constrained formulation (7) under certain conditions.

**Theorem 1.** Suppose that \( d(\cdot, \cdot) \) is the Wasserstein-1 distance, denoted by \( W_1(\cdot, \cdot) \), and \( \beta = 1 \). If \( \lambda > 1 \), then the optimal solution set of (8) is the same as that of (7).

**Proof.** To justify Theorem 1, we first consider the case of \( \lambda = 1 \), with which and \( \beta = 1 \), formulation (8) can be rewritten as
\[
\min_{f} \mathbb{E}\|Y - f(Y)\| + W_1(p_X, p_{f(Y)}).
\]

Let \(f^*\) be an optimal solution to (10) and \(L^*\) be the corresponding objective value. It is easy to see that \(f^*\) is also an optimal solution of
\[
\min_{f} \int \|Y - f(Y)\| dP_Y(y)
\]
subject to \(d(p_{f(Y)}; p_{f'(Y)}) = 0\),

which means \(\mathbb{E}\|Y - f*(Y)\| = W_1(p_Y, p_{f(Y)})\). Thus \(L^* = W_1(p_X, p_{f(Y)}) + W_1(p_Y, p_{f'(Y)})\).

Next, we show that \(L^* \geq W_1(p_X, p_Y)\). Suppose \(p_{X,Y}\) minimizing \(\int_{x,y} p_{X,Y}(x,y) ||X - Y|| dx dy\), then
\[
W_1(p_X, p_Y) = \int_{x,y} p_{X,Y}(x,y) ||x - y|| dx dy
\]
\[
= \int_{x,y,z} p_{X,Y,Z}(x,y,z) ||x - z - y|| dx dy
\]
\[
\leq \int_{x,y,z} p_{X,Y,Z}(x,y,z) (||x - z|| + ||z - y||) dx dy
\]
\[
= \int_{x,z} p_{X,Z}(x,z) ||x - z|| dx
\]
\[
+ \int_{y,z} p_{Y,Z}(y,z) ||z - y|| dy
\]
\[
= W_1(p_X, p_Z) + W_1(p_Y, p_Z).
\]

Let \(z = f^*(y)\), we have \(L^* \geq W_1(p_X, p_Y)\), which is obviously achievable when \(f(y) = y\). Thus \(L^* = W_1(p_X, p_Y)\).

When \(\lambda > 1\) and with \(\beta = 1\), (8) can be written as
\[
\min_{f} \mathbb{E}\|Y - f(Y)\| + \lambda W_1(p_X, p_{f(Y)})
\]
\[
\geq \min_{f} (\mathbb{E}\|Y - f(Y)\| + W_1(p_X, p_{f(Y)}))
\]
\[
+ (\lambda - 1) \min_{f} W_1(p_X, p_{f(Y)})
\]
\[
= L^* + (\lambda - 1) \min_{f} W_1(p_X, p_{f(Y)})
\]
\[
\geq W_1(p_Y, p_X).
\]

Inequality (13) gives the infimum of (8) when \(\lambda > 1\). Now, we prove that any optimal solution to (7) is also an optimal solution to (8). Let \(g^*\) be an optimal solution of (7), we have
\(\mathbb{E}\|Y - g^*(Y)\| + \lambda W_1(p_X, p_{g^*(Y)}) = W_1(p_Y, p_X)\), since \(p_{g(Y)} = p_X\) and \(W_1(p_X, p_{g(Y)}) = 0\). Then, it is easy to see from (13) that \(g^*\) is optimal to (8).

Last, we show that any optimal solution to (8) is also an optimal solution to (7). The equality in (13) holds only if \(W_1(p_X, p_{f(Y)}) = 0\), which means \(p_X = p_{f(Y)}\). Thus, minimizing (8) is equivalent to optimizing
\[
\min_{f} \mathbb{E}\|Y - f(Y)\| + \lambda W_1(p_X, p_{f(Y)})
\]
subject to \(p_X = p_{f(Y)}\),

which is equivalent to (7). Hence, the optimal solution set of (8) is the same as that of (7), which results in Theorem 1.

**Remark 1.** Interestingly, this theorem shows that, for \(d(\cdot, \cdot)\) being the Wasserstein-1 distance and \(\beta = 1\), the relaxed formulation (8) with any \(\lambda > 1\) has the same solution set as the constrained formulation (7). This result provides a solid theoretical foundation for the relaxation, which is used in practical implementation instead of (7). However, it is worth noting that, in practice the WGAN used to compute the Wasserstein-1 distance can only approximate the Wasserstein-1 distance to some extent [42]. As a consequence, to achieve satisfactory performance in practice, we still need to tune the parameter \(\lambda\) rather than selecting an arbitrary value of it.

### 4.2 An Information-Theoretic View of the Proposed Formulation

Here, from an information-theoretic view, we show the proposed formulation (7) has a favorable property that, the reconstruction map \(f(Y)\) can approximately achieve maximal preservation of the information of \(X\) contained in \(Y\).

First, the ideal supervised learning criterion is given by (2), which achieves perceptual reconstruction whilst maximizing the mutual information between \(X\) and \(\hat{X}\). In practical applications, mutual information is difficult to compute except for certain simple specific data distributions, hence MSE is widely used as the reconstruction loss, with which an implementation of the supervised criterion (2) can be written as
\[
\min_{f} \mathbb{E}_{(X,Y) \sim p_{X,Y}} (\|f(Y) - X\|^2)
\]
subject to \(d(p_X, p_{\hat{X}}) \leq 0\).

While general \(l_\beta\) loss with any \(\beta \geq 1\) can be alternative to MSE, here we consider MSE for convenience.

Under Gaussian distribution of \(p_X\) and \(p_Y\), \(\min_{f} \mathbb{E}_{Y \sim p_Y} (\|f(Y) - Y\|^2)\) is equivalent to maximizing the mutual information \(\max_{f} I(f(Y); Y)\) [43]. Hence, with \(\beta = 2\), the proposed unsupervised learning formulation (7) can be viewed as a special case of the following information-theoretic criterion for unsupervised learning under Gaussian distribution
\[
\max_{f} I(f(Y); Y)
\]
subject to \(d(p_X, p_{\hat{X}}) \leq 0\).

Next we show that the proposed unsupervised formulation is an approximation of the supervised formulation under certain conditions.

**Proposition 2.** Let \(X, Y, \hat{X}\) and \(N\) denote the source signal, degraded signal, restoration and noise, respectively, as defined in (1). Suppose that both \(X\) and \(\hat{X}\) are independent with \(N, p_X\) and \(p_Y\) are Gaussian distributed, then formulation (16) is equivalent to formulation (2), and formulation (7) with \(\beta = 2\) is equivalent to formulation (15).

**Proof:** First, under the constraint \(d(p_X, p_{\hat{X}}) \leq 0\), i.e. \(p_X = p_{\hat{X}}\), we have
\[
\mathbb{E} (||\hat{X} - X||^2) = \mathbb{E} ||\hat{X}||^2 + \mathbb{E} ||X||^2 - 2 \mathbb{E} (\hat{X}^T X)
\]
\[
= \text{const} - 2 \mathbb{E} (\hat{X}^T X).
\]
Similarly, it follows that
\[
\mathbb{E} (||\hat{X} - Y||^2)
\]
\[
= \mathbb{E} (||\hat{X} - X - N||^2) [\geq 2 \mathbb{E} ||X||^2 + \mathbb{E} ||N||^2 - 2 \mathbb{E} (\hat{X}^T X)
\]
\[
= \text{const} - 2 \mathbb{E} (\hat{X}^T X),
\]
where (a) is due to $d(p_X, p_{X'}) \leq 0$ and that both $X$ and $X'$ are independent with $N$. Thus, under the condition that both $X$ and $X'$ are independent with $N$, the unsupervised formulation (7) with $\beta = 2$ is equivalent to the supervised formulation (15).

Furthermore, under the assumption of Gaussian distribution of $p_X$ and $p_Y$, $\min_{f \sim p_{f \sim p_X}} \left( \| f(Y) - Y \|^2 \right)$ is equivalent to $\max_f I(f(Y); Y)$. Hence, it is easy to see the equivalence between the two information-theoretic formulations (16) and (2).

**Remark 2.** From the data processing inequality for the Markov chain $X \to Y \to X$, $I(\hat{X}; X)$ is upper bounded by $I(Y; X)$ as $I(\hat{X}; X) \leq I(Y; X)$. Naturally, if the restoration mapping $f$ can perfectly suppress the noise $N$ (hence $X$ is independent with $N$), maximally preserving the information of $X$ contained in $Y$ can be achieved by maximizing the mutual information $I(\hat{X}; Y)$. The assumption that the clean data $X$ is independent with the noise $N$ is reasonable in most applications. However, the assumption of independence between the reconstruction $\hat{X}$ and the noise $N$ is impractical, since the noise component in the observation $Y$ cannot be guaranteed to be completely suppressed. In practice, when the denoising process $f(Y)$ can suppress the noise component in $Y$ to a large extent, the correlation between $\hat{X}$ and $N$ would be weak. In this scenario, the unsupervised criterion (16) can be viewed as an approximation of the supervised criterion (2).

## 5 Experimental Results

We evaluate the proposed method on synthetic and real-world noisy images in comparison with state-of-the-art supervised, self-supervised, and unsupervised methods. The considered real-world images include photographic, microscopy, and raw depth images, for which the ground-truth is difficult or impractical to collect. The compared methods include DnCNN [1], RNAN [44], noise-to-clean (N2C), N2N [7], N2V [8], Nr2N [18], N2S [9], S2S [11], N2Same [45], NAC [19], BM3D [5], and CGAN (9). DnCNN, RNAN and N2C are supervised methods but use different network models. For fair comparison, N2C, N2N, CGAN and the proposed method use the same network model as detailed below in Section 5.2. The performance is evaluated in terms of both restoration distortion and perceptual quality measures, including peak signal to noise ratio (PSNR), structural similarity (SSIM), perception index (PI) [46] and learned perceptual image patch similarity (LPIPS) [47].

For real-world depth image denoising, the widely used colorization [48], cross-bilateral filter, nearest-neighbor hole-filling methods, and the Gaussian filter, median filter and bilateral filter denoising methods are also compared. Since the widely used Kinect camera only provide pre-processed images, we additionally conduct an experiment using another commercial time-of-flight (ToF) camera which can provide raw depth data without any pre-processing.

### 5.1 Network Settings

The proposed formulation (8) is implemented in an adversarial training framework using WGAN-gp [49], with $\beta = 1$ and $\lambda$ being tuned for each task. Note that although our formulation requires the access to clean images $X$, it is unsupervised since the clean images and the noisy ones are not required to be paired, which is similar to the unpaired image restoration task [50]. We use a U-Net architecture with skip connection for the generator, as shown in Fig. 2a. It consists of two down-sampling CNN layers in the encoder and two up-sampling CNN layers in the decoder. Residual channel attention block (RCAB) [51] is used for each down-sampling and up-sampling layer. The discriminator is the same as that in [35]. We remove the batch norm layers and sigmoid activation of the discriminator as [49] did. The architecture is shown in Fig. 2b. We train the generator and the discriminator alternatingly for a total of 200 epochs and use the RMSProp optimizer with a learning rate of $1 \times 10^{-4}$ for the generator and $0.5 \times 10^{-4}$ for the discriminator. The learning rate is decayed by a factor of 10 after 100 epochs. N2C, N2N, CGAN and the proposed method use the same restoration model in Fig. 2a.

For adversarial training in each task, we collect 20,000–60,000 patches of size $64 \times 64$ from clean images as the clean target domain, and the same number of patches of noisy images as the noisy source domain. For real-world photographic and microscopy images, the clean images are collected by averaging over tens of realizations of the same scenes, while clean depth images are generated by 3D modeling [52]. Moreover, ground-truth clean targets of noisy patches are collected to train N2C.

### 5.2 Synthetic Noisy RGB Images

The considered synthetic noise types include additive Gaussian noise, Poisson noise and Brown Gaussian noise.
For Gaussian noise, different noise levels with standard deviation $\sigma \in \{15, 25, 50\}$ are considered. The Brown Gaussian noise is generated by applying a Gaussian filter (with kernel size $5 \times 5$) to Gaussian noise with standard deviation of $\sigma = 50$. For Poisson noise, we set the maximum event count $\lambda_p = 30$ and the noise is implemented as $Y_i = \text{Poisson}(\lambda_p X_i)/\lambda_p$ where $i$ is the color channel.

The BSDS500 dataset [53] is used for training. Evaluation results are provided on the KODAK24 dataset [54], which contain 24 outdoor images. We train the N2C, N2N, and CGAN models using the same network as the proposed method. N2N is trained using noisy pairs with 2 realizations of each scene. N2V, N2S, S2S, N2Same, NAC are retrained using their original network models. The original DnCNN model trained in [1] is directly used as it is blind to noise. Nr2N [18] is only tested for Gaussian noise, as only its model trained in [1] is directly used as it is blind to noise. Table 1 shows the PSNR and SSIM results of the compared methods, whilst Table 2 shows the PI and LPIPS scores. PI and LPIPS are two perception quality measures that accord well with human perception. It can be seen that, N2C outperforms DnCNN, which is mainly due to more sophisticated network architecture using residual channel attention blocks. For spatially independent Gaussian and Poisson noise, the PSNR of our method is about 1 dB lower than the supervised method N2C. Our method outperforms BM3D in the cases of Poisson and Brown Gaussian noise, and the advantage is more conspicuous in Brown Gaussian noise. Particularly, for Brown Gaussian noise, which is spatially correlated, our method even has better performance than the supervised method N2C in terms of PSNR. Besides, it can be seen from Table 2 that our method achieves the best PI and LPIPS scores, which demonstrate that it can yield the best perceptual quality.

Fig. 3 compares the visual quality of typical restoration by the methods. It is clear that our method can achieve significant better perceptual quality than DnCNN, N2C, N2N,
N2V and BM3D, as it preserves more detail information while DnCNN, N2C, N2N and BM3D yield over-smoothed images. It is especially obvious when the image has abundant details. This advantage is due to the fact that the optimal transport criterion seeks a restoration that has the same distribution as natural clean images.

On the whole, in terms of both distortion (PSNR/SSIM) and perceptual quality (PI/LPIPS), our method even compares favorably with state-of-the-art supervised methods, especially in the case of complex noise, e.g., spatially correlated Brown Gaussian noise. The results also accord well with recent findings on the distortion-perception trade-off that, better perception quality can only be achieved at some increase of the distortion, e.g., MSE, PSNR and SSIM [55], [56].

Moreover, Table 3 shows the results of the proposed method for different values of \( \lambda \) in the condition of Gaussian noise with \( \sigma = 15 \). While Theorem 1 shows that any \( \lambda > 1 \) is an optimal selection, the empirical results show that a value of \( \lambda < 1 \) achieves the highest PSNR. This is because

![Visual comparison on synthetic noisy images with different noise conditions. The PSNR/PI/LPIPS results are provided in the brackets. The images are enlarged for clarity.](image-url)
in practice the used WGAN can only approximate the Wasserstein-1 distance.

### 5.3 Synthetic Noisy Depth Images

Depth images only capture the geometry of scenes, which is much different from RGB images. We consider the denoising of depth images with synthetic Gaussian noise with different standard deviation \( \sigma \in \{15, 25, 50\} \). The modes are trained on a simulated depth image dataset SUNCG [52], which contains 45,622 images of scenes with realistic room and furniture layouts.

Table 4 shows the PSNR and SSIM results of the compared methods on 27 images from the Middlebury dataset [57], [58]. Since the distribution of depth images is much simpler than that of color images, for the same level of noise each method can yield better PSNR and SSIM results than in the RGB experiments. It can be seen that, as the noise strength increases, the PSNR and SSIM of our method get closer to that of supervised methods. For example, in the case of noise strength \( \sigma = 50 \), the PSNR of N2C is 34.88 dB, while that of our method is 34.43 dB.

Fig. 4 compares the visual quality of the methods on a typical restoration example in the case of \( \sigma = 25 \). Again, our method yields better perceptual quality than the supervised methods and BM3D, as it preserves more details. The better perception quality is achieved at some necessary increase of PSNR [55], [56].

### 5.4 Real-world Microscopy Images

Next, we consider several real-world denoising tasks, where the ground-truth is difficult to collect. The first is microscopy image denoising. Microscopy images are an important source of data for biological and medical research. However, due to factors such as illumination and equipment during acquisition, microscopy images are inevitably corrupted by noise, which would affect subsequent high-precision analysis. Besides, since clean reference images are unavailable, unsupervised or self-supervised methods are desired.

We use a real fluorescence microscopy dataset, FMD [60], for training and testing. The dataset consists of 12,000 real fluorescence microscopy images of different samples obtained with different microscopes. It uses an image averaging strategy to collect pseudo ground-truth. Specifically, for each scene, 50 images are captured and averaged to get the pseudo ground-truth. N2C is trained with the supervision of this pseudo ground-truth. We use the noisy images

![Fig. 4. Visual comparison on synthetic noisy depth images in the case of Gaussian noise with \( \sigma = 25 \). The images are enlarged for clarity.](image-url)
in the training set to train CGAN, N2V, and our model. For N2N, we randomly select two from the 50 noisy images of each scene to generate training pairs. In this experiment, we consider an improved version of BM3D for microscopy images, VST+BM3D [59]. Two variants of our method are considered. The first uses the averaged images in FMD as clean source, denoted by “ours (pseudo)”. The second uses clean images from another microscopy image dataset collected by the approach of FairSIM [61], denoted by “ours (clean)”.

Table 5 shows the quantitative results on the test set of FMD. It is worth noting that all these results are computed based on the pseudo ground-truth. We also provide the PI and LPIPS scores to evaluate the perceptual quality of restored microscopy images by treating them as color images. It can be seen that our method achieves the highest PSNR, best PI and LPIPS scores, while N2C yields the highest SSIM. Even though using clean images from another source degrades the performance of our method a bit, it still gets the highest PSNR/SSIM among the compared self-supervised and unsupervised methods, and at the meantime achieves the best PI/LPIPS scores other than “ours (pseudo)”. The results demonstrate the feasibility and effectiveness of using a similar but different source of clean images in training our method. This means our method can be easily applied to the applications in which clean source images are unavailable.

Fig. 5 presents visual comparison on a typical restoration example. For visual clarity, we process the grayscale images of the R, G, B three channels separately, and then merge the three restored channels into a color image. Clearly, the restoration of our method is much clearer than that of VST+BM3D, N2C and N2N, which demonstrates its much better perceptual quality. More examples are provided in the supplemental material.

### 5.5 Real-world Photographic Images

This experiment evaluates the methods on denoising real-world photographic noisy images. In practical photography, due to factors such as ambient light and exposure time, the captured photos often have noise, which leads to unsatisfactory visual quality. Besides, because of poor aperture and photoreceptor used in mobile phone cameras, noise is common in the captured images.

The N2C, N2N, CGAN, N2V, N2S, S2S, N2Same, and our models are trained on a real photograph image dataset SIDD [63], which is collected by smartphone cameras and contains about 30,000 noisy images from 10 scenes under different lighting conditions. For each scene, the pseudo ground-truth is collected as follows: first continuously shoot...
multiple images of the scene, then impose image processing (such as image registration, abnormal image removal, etc.), and finally adopt weighted average to obtain the ground-truth of this scene. This pseudo ground-truth is used for training N2C. For N2N, we randomly select two from the noisy images of each scene to generate training pairs. Since the size of the test images is too large (e.g. 4000 x 3000), we crop the test images into patches of size 1024 x 1024 for testing. Similar to the microscopy image deoising experiment, two variants of our method are considered. The first uses the averaged images in SIDD as the clean source, denoted by “ours (pseudo)”. The second uses clean images from a different source collected by a HUAWEI Mate 40 smartphone, denoted by “ours (clean)”. 

Table 6 presents the quantitative results on the test set of SIDD. The ground-truth generated by post-processing and averaging is used to compute the PSNR and SSIM results. Accordingly, the results do not exactly reflect the difference from the real clean images. Our method achieves a PSNR of 39.69 dB, which is about 1.2 dB lower than that of N2C but about 1.7 dB higher than that of BM3D. Again, our method gets the best PI and LPIPS scores, which demonstrates that it can achieve the best perceptual quality. For our method, using clean images from another source leads to some degradation of performance, but it still outperforms all other self-supervised and unsupervised methods in PSNR/SSIM, and at the meantime outperforms all the supervised, self-supervised and unsupervised methods other than “ours (pseudo)” in terms of PI/LPIPS scores.

Fig. 6 presents visual comparison on a restoration example. Since the noise level is small and the restoration PSNRs of the methods are quite high, e.g., near 40 dB, the difference in visual quality is not distinct.

5.7 Real-world Depth Images

Recently, depth cameras have become increasingly popular, e.g., it is widely equipped in latest mobile phones and tablets. Due to its imaging mechanism, depth images have much larger noise than RGB images. Besides, there commonly exist holes (invalid pixels) in depth images due to low reflectivity and transparency of objects in the scene.

We first consider the denoising of Kinect depth images. As the Kinect camera is equipped with built-in filters, the output depth images have less noise but still have some holes. For training our model, we use the NYU dataset [62] as the noisy sample domain, and randomly collect the same amount of patches from the SUNCG dataset [52] as the clean target domain. The NYU dataset contains 1,449 depth images.

---

**Table 6:** Quantitative Comparison on Real-World Photographic Images

| Method      | PSNR/SSIM | PI/LPIPS |
|-------------|-----------|----------|
| Noisy       | 30.16/0.6349 | -        |
| Supervised |            |          |
| N2C         | 40.90/0.9651 | 8.53/0.033 |
| N2N [7]     | 40.88/0.9648 | 8.49/0.033 |
| Self-supervised | |         |
| N2V [8]     | 33.42/0.8226 | 7.87/0.149 |
| N2S [9]     | 32.98/0.7403 | 8.13/0.172 |
| S2S [11]    | 38.41/0.9459 | 7.55/0.035 |
| N2Same [45] | 35.81/0.9429 | 8.23/0.034 |
| Unsupervised |         |          |
| BM3D [5]    | 37.97/0.9495 | 8.45/0.032 |
| CGAN        | 38.20/0.9487 | 7.78/0.030 |
| Ours (clean)| 38.67/0.9502 | 7.53/0.027 |
| Ours (pseudo)| 39.69/0.9527 | 7.42/0.015 |

---

Fig. 6. Visual comparison on real-world photographic images, where the pseudo ground-truth is obtained by averaging over 10 realizations of each scene. The PSNR/PI/LPIPS results are provided in the brackets. The images are enlarged for clarity.
images of indoor scenes and the corresponding RGB images captured by a Kinect camera.

Since we have neither the corresponding ground-truth images nor noisy image pairs on the same scenes, the supervised methods N2C and N2N are not compared. We mainly compare the proposed method with representative and widely used inpainting methods for depth images, including the colorization, cross-bilateral filter, and nearest-neighbor methods. The classic nearest-neighbor method only uses a single depth image as input, whilst the colorization and cross-bilateral filter methods are RGB-D fusion methods, which take depth and corresponding RGB images as input.

Fig. 7 presents the restoration on a typical example. We colorize the depth image for visual clarity, in which the dark-red color represents invalid pixels with values of zero, i.e., holes. Obviously, our method achieves significantly better restoration quality than the nearest-neighbor method and the RGB-D fusion methods, colorization and cross-bilateral filter. The restoration of our method, which has straighter and sharper edges, is more consistent with the geometry of the scene (see the RGB of the scene). More examples are provided in the supplemental material.

5.8 Real-world Raw Depth Images

As the Kinect camera only provides pre-processed images, we additionally consider a raw depth denoising experiment using another commercial ToF camera which provides raw depth data without any pre-processing. We collect 1,430 raw depth images using this camera as the noisy sample domain. For the clean target domain, we also randomly collect the same amount of patches from the SUNCG dataset [52].

Since raw depth images have severe noise and invalid pixels, the denoising methods are used in combination with the nearest neighbor hole-filling methods. The compared denoising methods include N2V, BM3D, and three widely used methods for depth image denoising, Gaussian filter, median filter and bilateral filter. Our model can naturally learn to simultaneously fill holes and denoise at the same time. However, intensive extra experiments show that the following two-stage process can yield better performance than one-stage process. Specifically, since raw depth images contain many outliers and invalid pixels, we use the $\ell_1$ loss in (8) to train our model, which is robust to outliers. Then, as the restoration by this model still contain some noise, as shown in Fig. 8j, we cascade a post-denoiser to our model to obtain a refinement of the restoration. The evaluated post-denoisers include Gaussian filter, Bilateral filter, Median filter, N2V, BM3D and our model trained on synthetic depth images with Gaussian noise in Section 5.3.

Fig. 8 compares the restoration on a typical example. The depth images are colorized for visual clarity, in which the dark-red color represents zero-value invalid pixels. It can be seen from Fig. 8 that the raw depth image contains severe complex noise. Our method with a post-denoiser significantly outperforms the NN+BM3D, NN+N2V and the three filter based methods, as it can fill in the holes more naturally while preserving more edge details of the images. Particularly, the best performance is achieved when our model is paired with the denoiser trained on synthetic depth images with Gaussian noise. We provide more examples in the supplemental material.
A significant feature of the depth denoising task is that, it is generally impossible to collect ground-truth reference images for supervised training. The methods [64], [65] construct paired training data by first reconstructing the 3D geometry of the scene. But 3D geometry reconstruction usually suffers from errors. Besides, there exists another challenge of pixel-to-pixel alignment in constructing paired training data from 3D reconstruction. Our method shows remarkable superiority in denoising raw depth data but requires neither paired training data nor any prior information on the degradation model.

6 CONCLUSION

An unsupervised restoration learning criterion has been proposed from the optimal transport theory, which does not assume any prior knowledge on the degradation model. Favorable properties of this criterion have been revealed, e.g., approximately maximal preservation of the information of the signal, whilst achieving perceptual reconstruction. Moreover, for the relaxed formulation of this criterion, which is used in practical implementation, we proved that in theory the relaxation has the same solution as the original criterion. Experiment results on synthetic and real-world data demonstrated that the new method even compares favorably with supervised methods. Particularly, the results demonstrated remarkable superiority of the new method in denoising raw depth images with complex noise, which may be preferred in other practical applications with harsh circumstances in the wild but without any prior knowledge on the degradation model.

ACKNOWLEDGMENTS

Wei Wang, Fei Wen, and Zeyu Yan Contributed equally to this work.
REFERENCES

[1] 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., vol. 26, no. 20, pp. 3142–3154, Dec. 2017.

[2] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2802–2810.

[3] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image restoration,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[4] A. Bousquet, B. C. and J.-M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 60–65.

[5] K. Babov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” IEEE Trans. Image Process., vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[6] D. Ulyanov, V. Vedaldi, and V. Lempitsky, “Deep image prior,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 9446–9454.

[7] J. Lehtinen et al., “Noise2noise: Learning image restoration without clean data,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 4620–4631.

[8] A. Krull, T.-O. Buchholz, and F. Jüg, “Noise2void-learning denoising from single noisy images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2129–2137.

[9] J. Batson and L. Roter, “Noise2self: Blind denoising by self-supervised multi-scale self-supervised learning,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1465–1475.

[10] T. Ehret, A. Davy, J.-M. Morel, G. Facciolo, and P. Arias, “Model-blind video denoising via frame-to-frame training,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11369–11378.

[11] Y. Quan, M. Chen, T. Pang, and H. Ji, “Self2Self with dropout: Learning self-supervised denoising from single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1890–1898.

[12] M. Zhussip, S. Soltanayev, and S. Y. Chun, “Extending stein’s unbiased risk estimator to train deep denoisers with correlated pairs of noisy images,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 6970–6980, 2019.

[13] X. Wu, M. Liu, Y. Cao, D. Ren, and W. Zuo, “Unpaired learning of deep image denoising,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 204–219.

[14] A. Buades, B. Coll, and J.-M. Morel, “A review of image denoising algorithms, with a new one,” Multiscale Model. Simul., vol. 4, no. 2, pp. 490–530, 2005.

[15] F. Wen, R. Ying, Y. Liu, P. Liu, and T.-K. Truong, “A simple local minimal intensity prior and an improved algorithm for blind image deblurring,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 7, pp. 2099–2111, Jul. 2021.

[16] M. Lebrun, “An analysis and implementation of the bm3d image denoising method,” Image Process. On, vol. 2012, pp. 175–213, 2012.

[17] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, “Noiser2noise: Learning to denoise from unpaired noisy data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12064–12072.

[18] J. Xu et al., “Noisy-as-clean: learning self-supervised denoising from corrupted image,” IEEE Trans. Image Process., vol. 29, pp. 9316–9329, 2020.

[19] S. Cha, T. Park, B. Kim, J. Baek, and T. Moon, “GAN2GAN: Generative noise learning for blind denoising with single noisy images,” in Proc. Int. Conf. Learn. Represent., 2020.

[20] A. Krull, T. Vicar, M. Prakash, M. Lalit, and F. Jüg, “Probabilistic noise2void: Unsupervised content-aware denoising,” Front. Comput. Sci., vol. 2, 2020, Art. no. 5.

[21] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, “Neighbor2Neighbor: Self-supervised denoising from single noisy images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14781–14790.

[22] S. Soltanayev and S. Y. Chun, “Training deep learning based denoisers without ground truth data,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 3261–3271.

[23] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.

[24] G. Chryssos, J. Kossaifi, and S. Zafeiriou, “RoCGAN: Robust conditional GAN,” Int. J. Comput. Vis., vol. 128, no. 10, pp. 2665–2683, 2020.

[25] A. Bora, E. Price, and A. G. Dimakis, “Ambientgan: Generative models from lossy measurements,” in Proc. Int. Conf. Learn. Represent., 2018.
[53] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 898–916, May 2010.

[54] R. Franzen, “Kodak lossless true color image suite,” 1999. [Online]. Available: http://r0k.us/graphics/kodak/

[55] Y. Blau and T. Michaeli, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 675–685.

[56] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, “On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 11682–11692.

[57] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[58] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[59] M. Makitalo and A. Foi, “Optimal inversion of the generalized anscombe transformation for poisson-gaussian noise,” IEEE Trans. Image Process., vol. 22, no. 1, pp. 91–103, Jan. 2013.

[60] Y. Zhang et al., “A poisson-gaussian denoising dataset with real fluorescence microscopy images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11710–11718.

[61] M. Müller, V. Monkmöller, S. Hennig, W. Hübner, and T. Huser, “Open-source image reconstruction of super-resolution structured illumination microscopy data in imagej,” Nature Commun., vol. 7, no. 1, pp. 1–6, 2016.

[62] A. Janoch et al., “A category-level 3D object dataset: Putting the kinect to work” in Proc. Consum. Depth Cameras Comput. Vis., 2013, pp. 141–165.

[63] A. Abdelhamed, S. Lin, and M. S. Brown, “A high-quality denoising dataset for smartphone cameras,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1692–1700.

[64] Y. Zhang and T. Funkhouser, “Deep depth completion of a single RGB-D image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 175–185.

[65] J. Jeon and S. Lee, “Reconstruction-based pairwise depth dataset for depth image enhancement using cnn,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 422–438.

Wei Wang received the BS degree in computer science and technology from Shanghai Jiao Tong University, China, in 2020. He is currently working toward the master’s degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His main area of research interest is image processing.

Fei Wen (Senior Member, IEEE) received the BS degree from the University of Electronic Science and Technology of China (UESTC) in 2006, and the PhD degree in information and communications engineering from UESTC in 2013. Currently, he is a research associate professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His main research interests are signal/image processing and machine learning.

Zeyu Yan received the BEng degree from the University of Electronic Science and Technology of China (UESTC) in 2019. Since 2019, he has been with the Brain-Inspired Application Technology Center (BATC), Shanghai Jiao Tong University (SJTU), where he is currently working toward the PhD degree. His main areas of research interests are image processing and machine learning.

Peilin Liu (Member, IEEE) received the PhD degree from the University of Tokyo majoring in Electronic Engineering in 1998 and worked there as a research fellow in 1999. From 1999 to 2003, she worked as a senior researcher with the Central Research Institute of Fujitsu, Tokyo. Her research interests include low power computing architecture, application-oriented SoC design and verification, and 3D vision. She is currently a professor with Shanghai Jiao Tong University.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.