UNSUPERVISED CONTRASTIVE LEARNING FOR SIGNAL-DEPENDENT NOISE SYNTHESIS

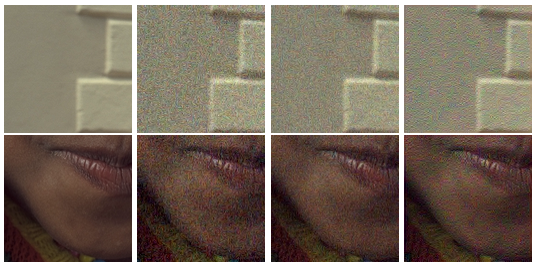
Anonymous authors
Paper under double-blind review

ABSTRACT

We present a simple yet robust noise synthesis framework based on unsupervised contrastive learning. With access to clean images only, the proposed contrastive noise synthesis framework trains a Glow-based generative model to synthesize image noise in a self-supervised fashion. We utilize the signal-dependency of the synthetic noise as a discriminative feature for the instance-wise discrimination pretext task and introduce a *noise contrastive loss* based on maximum mean discrepancy. The empirical results show that, with only 4312 parameters, the noise synthesized by the proposed framework shows advantages over the noise synthesized by traditional statistical models both qualitatively and quantitatively. The proposed framework fills a methodological gap in learning-based noise synthesis and can be used as an alternative to traditional statistical models.

1 INTRODUCTION

With a long-standing history in computer vision and image processing, image noise synthesis has been an important and active research topic. The additive synthetic noise has been widely utilized in applications such as data augmentation, noise modeling and reduction, and simulated robustness testing. Intuitively, the noise generation process can be understood as a one-to-many mapping due to the *stochasticity* of noise models, *i.e.* a clean image could have many noisy variants synthesized by the same noise model.



(a) Clean (b) AWGN (c) NLF (d) CoNo

Figure 1: Synthetic noisy images generated by (b) additive white Gaussian noise (AWGN), a homoscedastic Gaussian model; (c) noise level function (NLF), a signal-dependent heteroscedastic Gaussian model; and (d) CoNo, the proposed unsupervised contrastive learning framework. Compared with AWGN and NLF, CoNo does not have the inductive bias caused by the statistical independence between the pixels. CoNo generates noise with unnoticeable local patterns, which can be used to mimic the non-photon noise (*e.g.* fixed-pattern noise and spatially-correlated noise). The images are from the Kodak dataset (Malvar et al., 2004). Best viewed in color, with digital zoom.

A cheap solution is statistical modelling, where noise can be sampled from simple statistical distributions such as Gaussian or Poisson. However, recent studies (Plotz & Roth, 2017; Abdelhamed et al., 2018) argue that these statistical models cannot fully represent the characteristics of real noise. The statistical models simplify the real problem by imposing strong priors. For example, a homoscedastic Gaussian model ignores the signal-dependency of photon noise, *i.e.* the variance of the noise is proportional to the magnitude of the signal, and a heteroscedastic Gaussian model, *a.k.a.* the noise level function (NLF), can not model the non-photon noise (such as fixed-pattern noise and spatially-correlated noise) and non-linearities (such as amplification noise and quantization noise).

Fueled by the advance of deep learning, efforts have been made to learn realistic noise explicitly given large-scale noisy and clean image pairs as the training data. The state-of-the-art (SOTA) approaches (Chen et al., 2018; Abdelhamed et al., 2019; Chang et al., 2020) model the stochastic

noise generation process as a (conditional) generative model using GAN (Goodfellow et al., 2014) or Flow (Rezende & Mohamed, 2015). However, in most scenarios of noise synthesis, only clean images are available (Vincent et al., 2008), which makes the above noise learning approaches less practical. In addition, how to generate realistic signal-dependent noise via a self-supervised approach remains an open question.

Contributions To bridge the aforementioned methodological gap, we introduce a data-driven noise synthesis framework, named CoNo (stands for **Co**ntrastive **No**ise). CoNo integrates the concepts of unsupervised contrastive learning (UCL) (He et al., 2020; Chen et al., 2020), a powerful self-supervised learning (SSL) framework, and generative modeling. The key of UCL is to define an instance-wise discrimination *pretext* task, where the pretext task can be solved by minimizing a contrastive loss. The model of interest is then trained without any labels. We propose to use the signal-dependency of the synthetic noise as a discriminative feature in the self-supervised pretext task. To the best of our knowledge, this is the first effort of UCL in the domain of noise synthesis. While existing contrastive losses are designed to handle semantic feature vectors (Oord et al., 2018), we design a *noise contrastive loss* based on *maximum mean discrepancy* (Pan et al., 2010) that measures the distance between two distributions. CoNo consists of a *signal-dependent fusion* layer, which builds dependency between the homoscedastic Gaussian samples and clean images via *self-attention* (Vaswani et al., 2017), and a Glow-based (Kingma & Dhariwal, 2018) convolutional neural network (CNN). With only clean images, CoNo is able to generate realistic signal-dependent noise, as an alternative to traditional statistical models. The experimental results show that the noise generated by CoNo shows both qualitative and quantitative advantages over the noise generated by statistical models in the considered settings.

2 BACKGROUND

Statistical Noise Modeling Given a noisy image $\boldsymbol{y}$ and corresponding noise-free image $\boldsymbol{x}$, we have

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

where $\boldsymbol{n} \sim P(\boldsymbol{n})$ is the additive noise following an unknown distribution $P(\boldsymbol{n})$. The goal of statistical noise modeling or noise synthesis is to model $P(\boldsymbol{n})$.

The simplest model is to assume $\boldsymbol{n}$ is independent of $\boldsymbol{x}$, *i.e.* $P(\boldsymbol{n}|\boldsymbol{x}) = P(\boldsymbol{n})$. The most common noise model is to assume $\boldsymbol{n}$ follows a homoscedastic Gaussian distribution (Majumdar, 2018; Lehtinen et al., 2018; Batson & Royer, 2019; Krull et al., 2019; Guo et al., 2019; Yue et al., 2019), *a.k.a.* the additive white Gaussian noise (AWGN). Thus,

$$n_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where n_i is the noise at pixel i of the image $\boldsymbol{x}$ and σ is the standard deviation. However, AWGN ignores the fact the noise could be signal-dependent (Healey & Kondepudy, 1994; Gow et al., 2007; Liu et al., 2008; Foi et al., 2008; Hasinoff et al., 2010; Makitalo & Foi, 2012).

To take the signal dependency into consideration, a popular choice[1] is the heteroscedastic Gaussian model (Mohsen et al., 1975; Liu et al., 2014), *a.k.a.* the noise level function (NLF). NLF is defined as

$$n_i \sim \mathcal{N}(0, \lambda_{\text{shot}} x_i + \lambda_{\text{read}}), \tag{3}$$

where λ_{shot} and λ_{read} are two parameters controlling the variance (Brooks et al., 2019).

To sum up, the statistical noise modeling provides a simple yet robust solution to model $P(\boldsymbol{n}|\boldsymbol{x})$ and generate synthetic noise. It is worth mentioning that, with clean and noisy image pair $(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{n}$ can also be modeled via a data-driven approach, *e.g.* using a CNN to model $P(\boldsymbol{n}|\boldsymbol{x}, \boldsymbol{y})$ (Chen et al., 2018; Abdelhamed et al., 2019; Chang et al., 2020). We focus on the situation that only $\boldsymbol{x}$ is available in this work.

Contrastive Learning The theoretical breakthroughs in unsupervised contrastive learning (UCL) has fueled the recent successes in self-supervised learning (SSL) (Chen et al., 2020; He et al., 2020; Misra & Maaten, 2020; Tian et al., 2020; Chuang et al., 2020). Let $\boldsymbol{v}$ denote a feature vector extracted from an image patch of interest. For image classification tasks, a CNN backbone,

[1] See Appendix A.1 for the description of other statistical signal-dependent models.

e.g. ResNet (He et al., 2016), is commonly used as the encoder. Given an *query* image patch q and $(N+1)$ *key* image patches, a positive pair $(\boldsymbol{v}^q, \boldsymbol{v}^0)$ is defined as two patches taken from the same image and a negative pair $(\boldsymbol{v}^q, \boldsymbol{v}^{i>0})$ is defined as two patches taken from different images. Commonly, the image patches are generated by randomly cropping the stochastically augmented images. A common choice for the contrastive loss is InfoNCE (Oord et al., 2018), which is formulated as,

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(\boldsymbol{v}^q, \boldsymbol{v}^0)/\tau)}{\sum_{i=0}^{N} \exp(\text{sim}(\boldsymbol{v}^q, \boldsymbol{v}^i)/\tau)} \tag{4}$$

where τ is a temperature parameter and $\text{sim}(\cdot, \cdot)$ is the cosine similarity between two feature vectors. Mathematically, minimizing Eq. 4 is equivalent to maximize the mutual information shared between two views of the same image.

Note, Eq. 4 poses several constraints on the problem formulation of the downstream tasks of interest. First, high-level vision tasks containing semantic information benefit more from the encoded feature vectors than low-level vision tasks such as noise synthesis. This is determined by the nature of the instance-wise discrimination pretext task. Second, $\text{sim}(\cdot, \cdot)$ requires that two feature vectors are projected into the same high-dimensional feature space, *i.e.* there will be an element-wise correspondence between two feature vectors.

3 Contrastive Noise Learning

3.1 Problem Formulation

Following Sec. 2, we use $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{n}$, $\hat{\boldsymbol{n}}$ to denote clean image, noisy image, real noise, and synthetic noise, respectively. A dataset of clean images $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{M}$ is given as the training set. The goal is to generate realistic signal-dependent noise given $\mathcal{D}$ without access to $\{\boldsymbol{y}_i\}_{i=1}^{M}$ or $\{\boldsymbol{n}_i\}_{i=1}^{M}$.

For simplicity, we denote the dimension of the image as $d = H \times W \times C$ for a C-channel image with height H and width W. Let $g_{\boldsymbol{\theta}} : \mathbb{R}^{2d} \mapsto \mathbb{R}^d$, where g is a function parameterized by $\boldsymbol{\theta}$, *e.g.* a CNN. Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma \boldsymbol{I}^d)$ be a d-dimensional homoscedastic Gaussian sample[2], where σ is a parameter controlling the noise variance level (*e.g.* $\mathcal{N}$ is a standard normal distribution when $\sigma = 1$). We define the generated noise as

$$\hat{\boldsymbol{n}} = g_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{x}). \tag{5}$$

So, the learning outcome is to find an optimal set of parameters $\boldsymbol{\theta}$ that generates realistic $\hat{\boldsymbol{n}}$. Note, $\boldsymbol{z}$ is just an AWGN and $g_{\boldsymbol{\theta}}$ could be viewed as a transformation of $\boldsymbol{z}$ conditioning on $\boldsymbol{x}$.

3.2 Pretext Task

To utilize UCL on noise synthesis, we need to define an instance-wise discrimination pretext task as the first step. Let $P(\boldsymbol{n}|\boldsymbol{x})$ be the underlying but unknown signal-dependent noise distribution. We make the following assumption based on empirical observation:

Assumption 1 *The divergence between the noise distributions of two patches from the same image instance should statistically be smaller than the divergence between the noise distributions of two crops from two different image instances.*

As long as the assumption that $P(\boldsymbol{n}|\boldsymbol{x})$ is true, Asm. 1 should be valid. In addition to the fact that the patches from the same image instance share the same camera and acquisition setting, Asm. 1 can also be understood from the perspective of the self-similarity within the same image. Because of the potential self-similarity, the signal-dependent noise of two views of the same image instance should be somewhat correlated, which makes the divergence small. We expect the synthetic noise to share the same assumption as the real noise. Thus, we use the synthetic noise as the discriminative feature, analogous to feature vectors in high-level vision, for the pretext task,

3.3 Noise Contrastive Loss

Similar to previous UCL frameworks, the proposed framework consists of an instance-wise discrimination pretext task and a contrastive loss. The overall workflow is depicted in Fig. 2.

[2]We use $\boldsymbol{I}^d$ to denote the $d \times d$ identity matrix.

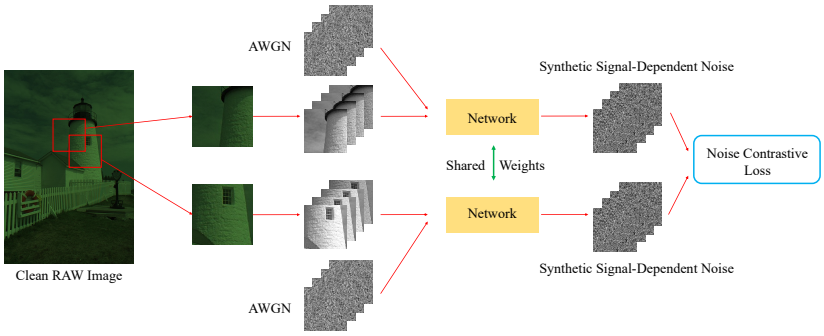


Figure 2: Illustration of the framework of CoNo. The random crops in shape $H \times W \times 1$ sampled from the RAW image (only a positive pair is shown) are rearranged into 4-channel (in order RGGB in this work) signal tensors in shape $H/2 \times W/2 \times 4$. Then, the signal tensors and additive white Gaussian noise (AWGN) samples are fed into the network to generate the noise for the instance-wise discrimination pretext task. The images displayed are proportionally scaled for better visualization.

In this work, we introduce a *noise contrastive loss* that is defined as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(-\text{dis}(\hat{\boldsymbol{n}}^q, \hat{\boldsymbol{n}}^0)/\tau)}{\sum_{i=0}^{N} \exp(-\text{dis}(\hat{\boldsymbol{n}}^q, \hat{\boldsymbol{n}}^i)/\tau)}, \tag{6}$$

where $\text{dis}(\cdot, \cdot)$ is a statistical distance. $\hat{\boldsymbol{n}}$ is the generated noise conditioning on the corresponding image patch $\hat{\boldsymbol{x}}$ cropped from the original image. Again, $(\hat{\boldsymbol{n}}^q, \hat{\boldsymbol{n}}^0)$ denotes the positive pair and $(\hat{\boldsymbol{n}}^q, \hat{\boldsymbol{n}}^{i>0})$ denotes the negative pair, given the query image patch $\hat{\boldsymbol{x}}^q$.

In contrast to Eq. 4, Eq. 6 replaces the statistical similarity with a negative statistical distance. We use $\text{dis}(\cdot, \cdot)$ to measure the divergence between two noise tensors instead of two feature vectors. Note, $\text{dis}(\cdot, \cdot)$ should be differentiable and be able to handle high-dimensional data. This is infeasible for most popular information theoretic approaches in generative modeling, such as *mutual information* (MI), *Kullback–Leibler divergence* (KLD), and *Jensen–Shannon divergence* (JSD) as the probability densities of two distributions have to be estimated beforehand.

Inspired by unsupervised domain adaptation (Pan et al., 2010), we consider *maximum mean discrepancy* (MMD) (Gretton et al., 2012) instead. It has been shown that MMD has several computational and statistical advantages over the above-mentioned distance measures (Smola et al., 2007)[3]. MMD is a nonparametric method based on the *kernel embedding of distributions* where a probability distribution is represented as an element of a *reproducing kernel Hilbert space* (RKHS)[4].

Given a domain Ω, let a function $f : \Omega \to \mathbb{R}$ belong to a class of functions $\mathcal{F}$. With $\hat{\boldsymbol{n}}^p = \{n_j^p\}_{j=1}^d$ conditioning on $\boldsymbol{x}^p$, we can view $\{n_j^p\}_{j=1}^d$ as samples from the probability distribution $P_p(\hat{\boldsymbol{n}}^p | \boldsymbol{x}^p)$. Similarly, we can define $\hat{\boldsymbol{n}}^q = \{n_j^q\}_{j=1}^d$ and $P_q(\hat{\boldsymbol{n}}^q | \boldsymbol{x}^q)$. We can define MMD and the empirical estimate of MMD (Borgwardt et al., 2006) between $(\hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q)$ respectively as

$$\text{MMD}(\mathcal{F}, P_p, P_q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_p[f(\boldsymbol{n}^p)] - \mathbb{E}_q[f(\boldsymbol{n}^q)]) \tag{7}$$

$$\text{MMD}(\mathcal{F}, \hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q) = \sup_{f \in \mathcal{F}} (\frac{1}{d} \sum_{j}^{d} f(n_j^p) - \frac{1}{d} \sum_{j}^{d} f(n_j^q)). \tag{8}$$

Given a Gaussian kernel k defined on $\Omega \times \Omega$

$$k(\omega, \omega') = \exp(-\frac{\|\omega - \omega'\|^2}{\delta^2}) \; \forall \omega \in \Omega \, \omega' \in \Omega, \tag{9}$$

[3]See Appendix A.2.1 for the advantages.
[4]See Appendix A.2.2 for the definition of RKHS.

Algorithm 1 Batch-wise training of *noise contrastive loss* for CoNo.

1: Sample a batch of $N+1$ images. ▷ Sample $N+1$ positive pairs
2: Sample two positive patches for each image.
3: Generate $\hat{\boldsymbol{n}}$ for each of $2N+2$ patches given σ. ▷ Eq. 5
4: **for** $j = 1, 2, \cdots, N+1$ **do** ▷ Compute *noise contrastive loss*
5: Take the j^{th} pair as the positive pair $(\hat{\boldsymbol{n}}^q, \hat{\boldsymbol{n}}^0)$.
6: Take the second patch of each of the other N pairs as $\hat{\boldsymbol{n}}^{i>0}$.
7: Compute $\mathcal{L}_{\text{contrast}}$ for the j^{th} positive pair. ▷ Eq. 6
8: Sum up $\mathcal{L}_{\text{contrast}}$ for a batch $N+1$ images as the batch-wise *noise contrastive loss*.

k is guaranteed to be universal. Now the statistical distance between $(\hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q)$ can be empirically computed[5] as

$$\mathrm{MMD}^2(\mathcal{F}, \hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q) = \frac{1}{d^2}\sum_{j}^{d}\sum_{j'}^{d} k(n_j^p, n_{j'}^p) + \frac{1}{d^2}\sum_{j}^{d}\sum_{j'}^{d} k(n_j^q, n_{j'}^q) - \frac{2}{d^2}\sum_{j}^{d}\sum_{j'}^{d} k(n_j^p, n_{j'}^q). \tag{10}$$

With Eq. 10, we can formally define Eq. 6 with $\mathrm{dis}(\cdot,\cdot) = \mathrm{MMD}^2(\cdot,\cdot)$. Specifically, we define two patches within the same image instance as a positive pair if they overlap. Two negative patches are defined as two patches without any overlap, either in the same image instance or in different image instances. This will provide the synthetic noise with a strong local similarity. Note, following the sampling process of UCL (Chen et al., 2020; He et al., 2020), we only sample positive pairs in the batch-wise training, illustrated in Algorithm 1.

It is worth mentioning that, while the workflow is similar to SimCLR (Chen et al., 2020), the proposed framework is not a trivial extension. CoNo differs from SimCLR in pretext task (noise tensors *vs.* feature vectors), downstream task (noise synthesis *vs.* semantic understanding), optimization (noise contrastive loss *vs.* InfoNCE), and modeling (a generative model *vs.* a discriminative model).

3.4 Network Architecture

$g_{\boldsymbol{\theta}}$ consists of two parameterized modules. The first module aims to fuse the AWGN and the clean image, *i.e.* to establish a signal-dependent prior distribution. The second module leverages the statistical power of flow-based models (Rezende & Mohamed, 2015; Kingma et al., 2016; Kingma & Dhariwal, 2018) in transforming a simple distributions into a complex one. For the second module, we just utilize Glow (Kingma & Dhariwal, 2018), a flow-based model comprising of a sequence of building blocks with identical architectures. Each building block has three layers, namely *actnorm*, *invertible* 1×1 *convolution*, and *affine coupling*.

Signal-Dependent Fusion As the first step in the noise generation process, we build a dependency between the AWGN $\boldsymbol{z}$ and the clean image $\boldsymbol{x}$, motivated by self-attention mechanism (Vaswani et al., 2017). We define a prior noise distribution as:

$$g_{\boldsymbol{\theta}_0}(\boldsymbol{z}, \boldsymbol{x}) = \sqrt{\boldsymbol{A} \otimes \boldsymbol{x} \oplus \boldsymbol{B}} \otimes \boldsymbol{z}, \;\; \boldsymbol{A} = \exp(a(\boldsymbol{x})), \;\; \boldsymbol{B} = \exp(b(\boldsymbol{x})), \tag{11}$$

where $g_{\boldsymbol{\theta}_0}$ implies that Eq. 11 is the zeroth step (the first step in the context of computer science), $\otimes$ and $\oplus$ denote element-wise multiplication and addition, respectively, and $a(\cdot)$ and $b(\cdot)$ are two shallow networks consisting of convolutional layers. Note, $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-negative and are dependent on $\boldsymbol{x}$. $\boldsymbol{A} \otimes \boldsymbol{x} \oplus \boldsymbol{B}$ is equivalent to simulate a signal-dependent variance, similar to Eq. 3. Intuitively, $g_{\boldsymbol{\theta}_0}(\boldsymbol{z}, \boldsymbol{x})$ scales up $\boldsymbol{z}$ by a factor of $\sqrt{\boldsymbol{A} \otimes \boldsymbol{x} \oplus \boldsymbol{B}}$ in an element-wise fashion. The signal-dependent fusion (SDF) layer, together with the following Glow architecture, can transform a simple Gaussian distribution $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}^d)$ into a complex data-dependent distribution.

3.5 Optimization

Note, the noise contrastive loss is only designed to enforce Asm. 1, trivial solutions exist, *i.e.* the synthesized noise is unrealistic. Besides, we often expect the noise generation process to be control-

[5] See Appendix A.2.3 for the derivation of $\mathrm{MMD}^2(\mathcal{F}, \hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q)$.

lable, like statistical models, in practical applications. In addition to the noise contrastive loss, we include two regularization terms below to ensure that CoNo can generate realistic noise.

Perceptual Loss As the noisy images and clean images should convey the similar underlying semantic information, we set an optimization objective to maintain the semantic information contained in the noisy image. To do so, we include a *perceptual loss* (Johnson et al., 2016) based on feature extracted from a pre-trained encoder f_e:

$$\mathcal{L}_{\text{perceptual}} = \|\phi(f_e(\boldsymbol{x} + \hat{\boldsymbol{n}})) - \phi(f_e(\boldsymbol{x}))\|_2^2, \tag{12}$$

where $\phi(\cdot)$ represents the features extracted from different layers of f_e.

Distributional Alignment Loss The unsupervised learning process can be further regularized if additional knowledge of $P(\boldsymbol{n}|\boldsymbol{x})$ is given. Commonly, when using synthetic noise to assess the performance of denoising algorithms (Majumdar, 2018; Lehtinen et al., 2018; Batson & Royer, 2019; Krull et al., 2019), the statistics of the noise for the target images are given for a quantitative comparison. For example, we always assume the noise has a zero mean and the denoising methods are usually evaluated under various noise variance levels. Thus, to make the synthetic noise controllable at different noise variance level, we want to align the learned distribution $P_{\boldsymbol{\theta}}$ with a prior statistical distribution P_{prior}. We implement this alignment via *moment matching*, *i.e.* we minimize the distance between the moments of $P_{\boldsymbol{\theta}}$ and P_{prior}. For the generated noise $\hat{\boldsymbol{n}}$ of image $\boldsymbol{x}$, the first moment (mean) and the second moment (variance) are

$$\mu(\hat{\boldsymbol{n}}) = \frac{\sum_j n_j}{d},\ \sigma^2(\hat{\boldsymbol{n}}) = \frac{\sum_j (n_j - \mu(\hat{\boldsymbol{n}}))^2}{d-1}. \tag{13}$$

Given $P_{\text{prior}} = \mathcal{N}(0, \sigma^2)$, the *distributional alignment loss* is

$$\mathcal{L}_{\text{alignment}} = \|\mu(\hat{\boldsymbol{n}})\|_2^2 + \|\sigma^2(\hat{\boldsymbol{n}}) - \sigma^2\|_2^2. \tag{14}$$

Final Objective The total loss is then the sum of three losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{perceptual}} + \mathcal{L}_{\text{alignment}}. \tag{15}$$

We use equal weights for three losses in this work, but different weights could be assigned.

4 Experiments

4.1 Experimental Setup

Datasets We use three public benchmark datasets: Kodak Image Dataset (Malvar et al., 2004), Darmstadt Noise Dataset (DND) (Plotz & Roth, 2017) and Smartphone Image Denoising Dataset (SIDD) (Abdelhamed et al., 2018). Kodak contains 25 images in PNG format with fixed resolution 768×512. Kodak has been widely adopted as the denoising dataset by adding the synthetic noise generated by statistical models described in Sec. 2. We use Kodak as the training set and visually analyze the noised generated by CoNo. DND contains 50 high-resolution images with realistic noise from 50 scenes taken by 4 cameras. SIDD contains thousands of noisy and clean image pairs, captured with five different smartphone cameras repeatedly from ten different scenes. As DND and SIDD contain real signal-dependent noise, we use DND and SIDD to illustrate the advantage of CoNo over statistical models quantitatively.

Implementation We focus on RAW images as they directly represent the noise distribution (Abdelhamed et al., 2019), while the rendering process from RAW images to RGB images can significantly change the noise distribution (Nam et al., 2016). So, we generate noise in RAW color space. For consistency, we set the Bayer color filter array pattern as `RGGB`. The input RAW crop has a size of $128\times128\times1$ and is reshaped into $64\times64\times4$ following (Gharbi et al., 2016). The RAW images are rendered into RGB images through a color processing pipeline (Menon et al., 2006).

Network CoNo consists of a SDF layer and a sequence of 8 identical building blocks of Glow (Kingma & Dhariwal, 2018), where each block consists of three layers: *actnorm*, *invertible* 1×1 *convolution*, and *affine coupling*. The sub-networks $a(\cdot)$ and $b(\cdot)$ in SDF both have three convolutional layers, followed by ReLU (Nair & Hinton, 2010) and batch normalization (Ioffe &

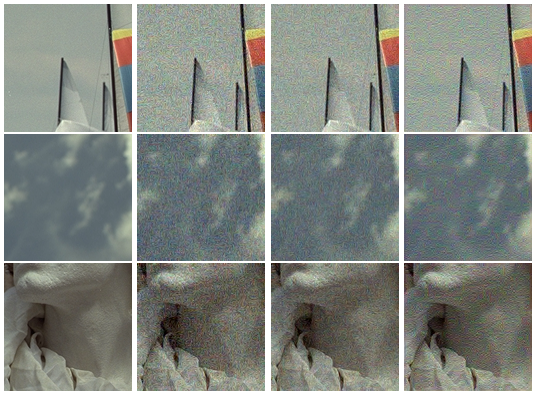
Szegedy, 2015). The input and output dimension pair for three layers are (4, 8), (8, 8), and (8, 4), respectively. The kernel size of three convolutional layers are 1×1, 3×3 and 1×1. Note, CoNo is a fairly simple model with only 4312 parameters.

Training We follow (Chen et al., 2020) in defining temperature $\tau = 0.07$ in Eq. 6. We use an Adam (Kingma & Ba, 2015) optimizer with a batch size of 64 and a fixed learning rate 0.03. Given the overall noise variance level σ^2 in Eq. 14, the network is trained for 50 epochs and we choose the best model parameters θ by selecting the epoch which provides the lowest $\mathcal{L}_{\text{total}}$ in Eq. 15.

Baselines Without access to real noise or any camera-calibrated parameters, we compare CoNo against two well-known baseline models, AWGN (Eq. 2) and NLF (Eq. 3). For a fair comparison, we aim to evaluate three models under the same overall variance level σ^2. Given σ, λ_{shot} and λ_{read} are sampled to ensure that the overall variance level is around σ^2.

4.2 Qualitative Analysis

Noise Synthesis In order to qualitatively analyze the synthetic noisy images, we first visualize the synthetic noisy images at the same noise variance level. Fig. 3 shows the generated noisy Kodak images for AWGN, NLF, and CoNo at noise variance level $\sigma = 10$. Note, although statistical models may describe photon noise, the real images have other noise sources (*e.g.* fixed-pattern noise, defective pixels, clipped intensities, spatially correlated noise, amplification noise, and quantization noise). These non-photon noise can not be easily modeled via statistical models. Moreover, statistical models such as AWGN and NLF tend to assume independence between the pixels. As a comparison, CoNo does not have this inductive bias. Meanwhile, the convolution operations give CoNo more flexibility in synthesizing the signal-dependent noise. If we take a closer look at Fig. 3, CoNo generates noise with unnoticeable local patterns, which simulate the non-photon noise and logically make the synthetic noisy images more similar to the real noisy images than rigid statistical models. In our experiment, independent human viewers suggest that the noisy images synthesized by CoNo are more perceptually comfortable than the noisy images synthesized by AWGN and NLF under the same noise variance level. We hypothesize that this is due to the fact that the noisy images generated by CoNo share more similar characteristics with the commonly seen real noisy images, such as intensity-based local patterns and pixel-wise dependence[6].



(a) Clean (b) AWGN (c) NLF (d) CoNo

Figure 3: Synthesized noisy images at the same noise variance level ($\sigma = 10$): (a) clean images: (b) AWGN, (c) NLF, (d) CoNo. CoNo shows more local noise patterns than the other two models.

Impact of Noise Variance Level To further understand the noise synthesized by CoNo, we examine the noise at different noise variance level (represented by the standard deviation σ).The noisy images generated by AWGN, NLF, and CoNo at $\sigma \in \{10, 20, 30\}$ are displayed in Fig. 4. With increased σ, CoNo tends to show more local patterns than NLF, while Gaussian has no patterns at all. Another interesting observation is that, given the same σ, UCL tends to show more artifacts in the bright regions than AWGN and NLF (see the white region of the parrot's beak). This phenomenon might be explained by the physical model of camera noise (Hasinoff et al., 2010)[7]: the bright regions in the photo have larger noise variances than the dark regions. The statistical models describe the signal-dependency variance via a linear relationship (*e.g.* Eq. 3). As a comparison, the data-driven approach can model more complex relationships.

[6]We conjecture that such a similarity may activate the human memory mechanism, thus make the viewer perceptually comfortable. The psychological discussion is beyond the scope of this work.

[7]See Appendix A.3 for the description of the physical model.

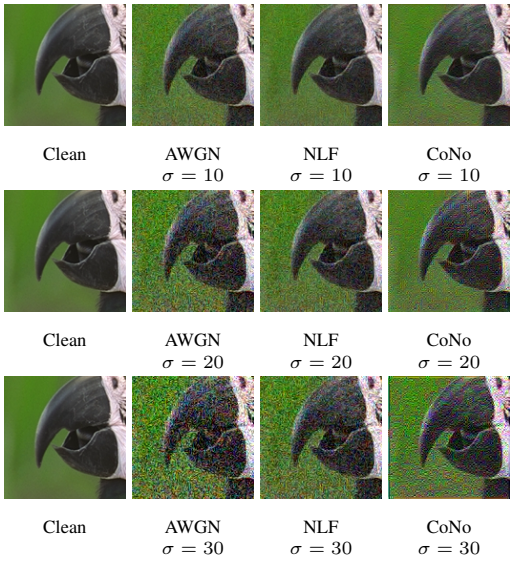


Figure 4: Generated noisy images at different noise variance level ($\sigma \in \{10, 20, 30\}$). In contrast to AWGN and NLF, CoNo tends to generate larger noise in the bright regions.

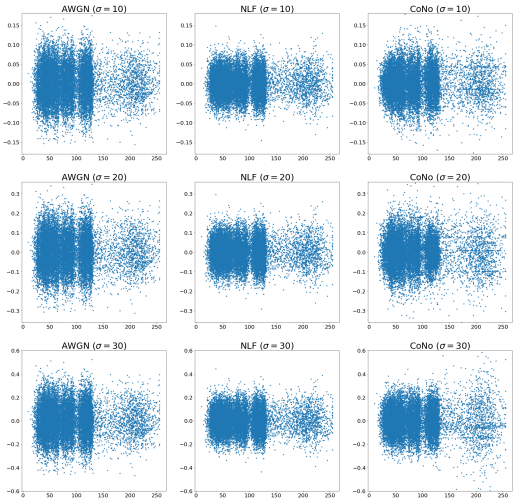


Figure 5: Scatter plots of pixel intensity and noise at different noise variance level ($\sigma \in \{10, 20, 30\}$) for Fig. 4. The horizontal axis represents the pixel intensity ($[0, 255]$). The vertical axis represents the noise at each pixel (normalized by 255). In contrast to AWGN and NLF, the noise generated by CoNo tends to show larger variance for high pixel intensity.

Analysis of Noise Distribution In addition to the qualitative comparison, we visualize the noise distribution to further validate the advantages of CoNo. We leverage statistical plots to study the relationship between pixel intensity $\boldsymbol{x}$ and noise $\hat{\boldsymbol{n}}$ in signal-dependent noise modeling. The scatter plots of pixel intensity and noise for three noise models are displayed in Fig. 5. The noise generated by CoNo shows similar patterns as the noise generated by AWGN and NLF for pixels with low intensity. However, for pixels with high intensity, the noise generated by CoNo tends to show larger variance for high pixel intensity than AWGN and NLF. This observation fits the physical phenomenon previously discussed, which further validates the hypothesis above: CoNo shows more flexibility in noise modeling than AWGN and NLF.

Noise Stochasticity As a generative model, a key property is the stochasticity of the noise synthesis, *i.e.* we expect the model can generate different noise given the same input image. Theoretically, the stochasticity comes from $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma \boldsymbol{I}^d)$ in Eq. 11. We visualize the noisy images generated by CoNo given different $\boldsymbol{z}$ in Fig. 6. The noisy images are slightly different.

4.3 Quantitative Analysis

The synthetic noise has been widely utilized in the applications of computer vision and image processing. The purposes of the experiments in this section are twofold. First, we notice that the real noise should be unavailable under our problem formulation. Thus, we use image denoising, a closely related problem to noise synthesis, as a *proxy* task to evaluate the quality of the synthetic

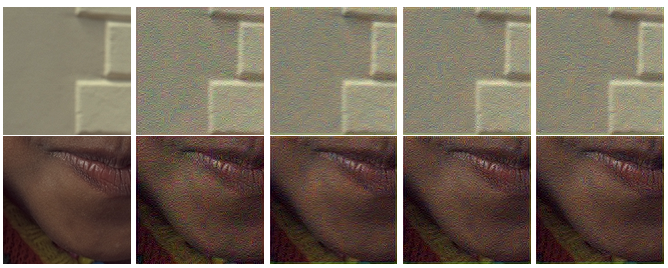


Figure 6: Stochasticity of the noisy images generated by CoNo. At each row, the first image is the clean image and the other four images are the generated noisy images given the same clean image but different random sample $\boldsymbol{z}$ ($\sigma = 10$ for the displayed images).

Table 1: DnCNNs are trained on the synthetic Kodak datasets and evaluated on DND and SIDD datasets, respectively. CoNo outperforms AWGN and NLF in the proxy evaluation.

	DND		SIDD	
Model	PSNR	SSIM	PSNR	SSIM
DN-AWGN	31.53	0.750	27.57	0.668
DN-NLF	31.96	0.759	28.68	0.685
DN-CoNo	**32.51**	**0.804**	**30.82**	**0.701**

Table 2: Under data scarcity (*e.g.* with only 10% of training set), there is a performance gain by pre-training DnCNN on the synthetic dataset. *Oracle* is trained with the full training set.

	DND		SIDD	
Model	PSNR	SSIM	PSNR	SSIM
w/o CoNo	34.07	0.851	32.59	0.861
w/ CoNo	**34.67**	**0.865**	**33.28**	**0.867**
Oracle	38.08	0.936	38.41	0.909

noise generated by CoNo
in a quantitative manner. Second, we illustrate that CoNo can be used as an alternative noise synthesis method in practical applications.

Proxy Evaluation Setup We create three large labeled training sets by generating 10000 noisy-clean RGB image pairs by AWGN, NLF, and CoNo. Specifically, for a clean image patch randomly cropped from Kodak dataset, we generate a noisy image given a randomly sampled σ, where $\sigma \sim \mathcal{U}(0, 30)$ and $\mathcal{U}$ is a uniform distribution. Note, we assume that the distribution of the test set is unknown during the training, *i.e.* we have no prior knowledge on the test set. Thus, we sample σ to cover possible variance levels as a common practice in denoising (Gharbi et al., 2016). For computational efficiency, we do batch-wise sampling for σ, while the other training details are the same as above. In this way, CoNo is trained to be sensitive to σ. We use DnCNN (Zhang et al., 2017), a SOTA image denoiser, as the backbone. We train three DnCNNs initialized with the same random seed on three training sets respectively in a standard supervised fashion (Abdelhamed et al., 2019), where the loss is $L1$. Then, we get three trained DnCNNs, which are denoted as DN-AWGN, DN-NLF, and DN-CoNo. We evaluate the trained DnCNNs by reporting the denoising performance on DND and SIDD, with metrics PSNR and SSIM.

Empirical Results In the first scenario, we directly evaluate the trained denoiser on the testing sets of DND and SIDD, which are unseen in the training. The results are presented in Table 1. Because DnCNNs are trained and evaluated in different datasets, *i.e.* the noise distributions are different, the results in Table 1 are only used as a proxy evaluation on the quality of the synthetic noise. As DN-CoNo outperforms DN-AWGN and DN-NLF by a large margin, we conclude that CoNo can synthesize more realistic noise than AWGN and NLF. As an ablation study, we demonstrate how CoNo could be used to in practical applications to mitigate the data scarcity. Two DnCNNs initialized with the same random seed are trained in parallel. One is trained under a standard supervised learning with 10% of the training set of the target tasks. Another one is firstly pre-trained with the synthetic dataset and then trained as the first DnCNN. As shown in Table 2, pre-training DnCNN on the synthetic dataset generated by CoNo does improve the performance.

4.4 Limitations

Although a UCL framework is proposed to fill a methodological gap in self-supervised noise synthesis, we should admit the limitations of CoNo: there is a trade-off between the performance and cost. Compared with statistical models, the training and inference phases of CoNo both require non-trivial computational cost and memory footprint. Meanwhile, the focus of experimental design is to evaluate the quality of the synthetic noise. The study on the related downstream tasks such as data augmentation and denoising is beyond the scope of the discussion, and is left as future work.

5 Conclusions

We are the first to propose an unsupervised signal-dependent noise synthesis framework based on contrastive learning. The proposed framework is simple yet robust and can be utilized as an alternative to traditional statistical models.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2018.

Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3165–3173, 2019.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11036–11045, 2019.

Ke-Chi Chang, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang, and Hwann-Tzong Chen. Learning camera-aware noise models. In *European Conference on Computer Vision*, pp. 343–358. Springer, 2020.

Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Alessandro Foi. Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing*, 89(12):2609–2629, 2009.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008.

Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Ryan D. Gow, David Renshaw, Keith Findlater, Lindsay Grant, Stuart J. McLeod, John Hart, and Robert L. Nicol. A comprehensive tool for modeling cmos image-sensor-noise performance. *IEEE Transactions on Electron Devices*, 54(6):1321–1329, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1722, 2019.

Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560. IEEE, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

G.E. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456. PMLR, 2015.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pp. 2965–2974. PMLR, 2018.

Ce Liu, Richard Szeliski, Sing Bing Kang, C. Lawrence Zitnick, and William T. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):299–314, 2008.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Practical signal-dependent noise parameter estimation from a single noisy image. *IEEE Transactions on Image Processing*, 23(10):4361–4371, 2014.

Angshul Majumdar. Blind denoising autoencoder. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1):312–317, 2018.

Markku Makitalo and Alessandro Foi. Optimal inversion of the generalized anscombe transformation for poisson-gaussian noise. *IEEE transactions on image processing*, 22(1):91–103, 2012.

Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pp. iii–485. IEEE, 2004.

Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE Transactions on Image Processing*, 16(1):132–141, 2006.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

A.M. Mohsen, M.F. Tompsett, and C.H. Sequin. Noise measurements in charge-coupled devices. *IEEE Transactions on Electron Devices*, 22(5):209–218, 1975.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1683–1691, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

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

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538, 2015.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pp. 1096–1103, 2008.

Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. In *Advances in Neural Information Processing Systems*, pp. 1690–1701, 2019.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017.

A Appendix

A.1 Signal-Dependent Models

In low-level computer vision, an early signal-dependent model is the Poisson model:

$$n_i \sim \alpha \mathcal{P}(x_i) - x_i, \tag{16}$$

where $\mathcal{P}$ is the Poisson distribution with rate x_i, x_i is the noise-free signal at pixel i of the image $\boldsymbol{x}$, and α is a scaling factor.

As there could be both signal-dependent and signal-independent components in the sources of noise, a stand-alone Poisson model is inefficient to fully explain noise. A more popular choice is to combine Eq. 2 and Eq. 16, which represent the signal-independent and signal-dependent parts respectively. The Poisson-Gaussian model (Foi et al., 2008; Foi, 2009; Makitalo & Foi, 2012) is

$$n_i \sim \alpha \mathcal{P}(x_i) - x_i + \mathcal{N}(0, \sigma^2). \tag{17}$$

A.2 MAXIMUM MEAN DISCREPANCY

A.2.1 ADVANTAGES

The advantages of MMD over information theoretic approaches can be summarized as below.

1. There are no restrictive assumptions about the form of the distributions and relationships between variables.
2. No intermediate density estimation is required.
3. The prior knowledge could be incorporated via choice of the kernel.
4. There is no information loss as the kernel embedding can uniquely preserve all information about a distribution.
5. It is fairly easy to compute.
6. It has good generalization ability.

A.2.2 DEFINITION OF RKHS

The *Moore–Aronszajn theorem* (Aronszajn, 1950) asserts the existence of RKHS $\mathcal{H}$ as a complete inner product space of functions $f : \Omega \mapsto \mathbb{R}$ with formal definitions of inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norms $\| \cdot \|_{\mathcal{H}}$. Let ϕ denote a *feature space map* $\Omega \mapsto \mathcal{H}$, the reproducing property holds:

$$\forall \omega \in \Omega \quad \langle f, \phi(\omega) \rangle_{\mathcal{H}} = f(\omega). \tag{18}$$

A.2.3 DERIVATION

Following (Borgwardt et al., 2006), we constrain $\|f \leq 1\|_{\mathcal{H}} \; \forall f \in \mathcal{F}$. Eq. 7 can be rewritten as

$$\begin{aligned}
\mathrm{MMD}(\mathcal{F}, P_p, P_q) &= \sup_{\|f \leq 1\|_{\mathcal{H}}} (\mathbb{E}_q[f(n^p)] - \mathbb{E}_p[f(n^q)]) && (19)\\
&= \sup_{\|f \leq 1\|_{\mathcal{H}}} (\mathbb{E}_p[\langle f, \phi(n^p) \rangle_{\mathcal{H}}] - \mathbb{E}_q[\langle f, \phi(n^q) \rangle_{\mathcal{H}}]) && (20)\\
&= \sup_{\|f \leq 1\|_{\mathcal{H}}} \langle f, \mu_p - \mu_q \rangle_{\mathcal{H}} && (21)\\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}, && (22)
\end{aligned}$$

where $\mu_p = \mathbb{E}_p[\phi(n^p)]$ and $\mu_q = \mathbb{E}_q[\phi(n^q)]$ are the expectation of ϕ in feature space. Eq. 20 is an application of the reproducing property in Eq. 18. Eq. 21 utilizes the linearity of the inner product. Eq. 22 utilizes the *Cauchy–Schwarz inequality*.

$$\begin{aligned}
\mathrm{MMD}^2(\mathcal{F}, \hat{\boldsymbol{n}}^p, \hat{\boldsymbol{n}}^q) &= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_{\mathcal{H}} \\
&= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\
&= \mathbb{E}_p[\langle \phi(n^p), \phi(n^p) \rangle_{\mathcal{H}}] + \mathbb{E}_q[\langle \phi(n^q), \phi(n^q) \rangle_{\mathcal{H}}] \\
&\quad - 2\mathbb{E}_{p,q}[\langle \phi(n^p), \phi(n^q) \rangle_{\mathcal{H}}]
\end{aligned} \tag{23}$$

Eq. 10 is just an empirical estimated of the last equation.

A.3 PHYSICAL CAMERA NOISE MODEL

A widely recognized physical camera noise model (Hasinoff et al., 2010) points out the variance of the signal-dependent noise can be modeled as

$$\sigma^2 = \frac{\Phi t}{g^2} + \frac{\sigma^2_{\mathrm{read}}}{g^2} + \sigma^2_{\mathrm{ADC}}. \tag{24}$$

The first term denotes the photon noise, where Φ is the radiant power, t is the exposure time, and g is the sensor gain. The terms σ^2_{read} and σ^2_{ADC} are the variance of the readout noise and the variance of the analog-to-digital conversion (ADC) noise.

A.4 DATASETS

We use three public benchmark datasets: Kodak Image Dataset (Malvar et al., 2004)[8], Darmstadt Noise Dataset (DND) (Plotz & Roth, 2017)[9] and Smartphone Image Denoising Dataset (SIDD) (Abdelhamed et al., 2018)[10]. For DND and SIDD, the datasets are randomly split into a training set with 80% of images and a testing set with 20% of images. We perform the denoising in the RGB color space.

[8] http://www.cs.albany.edu/~xypan/research/snr/Kodak.html

[9] https://noise.visinf.tu-darmstadt.de

[10] https://www.eecs.yorku.ca/~kamel/sidd/dataset.php