Joint Demosaicking and Denoising in the Wild:
The Case of Training Under Ground Truth Uncertainty

Jierun Chen, Song Wen, S.-H. Gary Chan
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong, China
{jcheneh, swenaa, gchan}@cse.ust.hk

Abstract
Image demosaicking and denoising are the two key fundamental steps in digital camera pipelines, aiming to reconstruct clean color images from noisy luminance readings. In this paper, we propose and study Wild-JDD, a novel learning framework for joint demosaicking and denoising in the wild. In contrast to previous works which generally assume the ground truth of training data is a perfect reflection of the reality, we consider here the more common imperfect case of ground truth uncertainty in the wild. We first illustrate its manifestation as various kinds of artifacts including zipper effect, color moire and residual noise. Then we formulate a two-stage data degradation process to capture such ground truth uncertainty, where a conjugate prior distribution is imposed upon a base distribution. After that, we derive an evidence lower bound (ELBO) loss to train a neural network that approximates the parameters of the conjugate prior distribution conditioned on the degraded input. Finally, to further enhance the performance for out-of-distribution input, we design a simple but effective fine-tuning strategy by taking the input as a weakly informative prior. Taking into account ground truth uncertainty, Wild-JDD enjoys good interpretability during optimization. Extensive experiments validate that it outperforms state-of-the-art schemes on joint demosaicking and denoising tasks on both synthetic and realistic raw datasets.

Introduction
Modern digital cameras use a single sensor overlaid with a color filter array (CFA) to capture an image. This means that only one color channel’s value is recorded for each pixel location. Let $N$ be the number of pixels in an image, the raw data acquisition process can be simply modeled as

$$x = Az + n,$$  

where $x \in \mathbb{R}^N$ is a noisy raw data vector of luminance readings, $A \in \mathbb{R}^{N \times 3N}$ is a mosaicking operation, $z \in \mathbb{R}^{3N}$ is an unknown clean image with three color channels, and $n \in \mathbb{R}^N$ is a noise vector.

Before the final “cooked” image is ready for the users, the raw data undergoes a series of processing steps, known as the image processing pipeline. Among those, demosaicking and denoising (DM&DN) are two of the very early and most crucial steps. Demosaicking aims to undo the mosaicking operation $A$ by interpolating the missing two-thirds of each pixel’s color channels, while denoising removes the inevitable noise $n$ from the measurement $x$. Due to their mod- ular property, substantial traditional literature takes them as independent tasks and executes them in a sequential manner. This yields potentially suboptimal performance, and inspires several works on jointly addressing the DM&DN tasks (Liu et al. 2020; Kokkinos and Lefkimmiatis 2019; Tan et al. 2017a).

Among the joint DM&DN works, data-driven approaches (Liu et al. 2020; Tan et al. 2018; Kokkinos and Lefkimmiatis 2018) have been shown more effective than applying handcrafted priors and filters. These approaches usually require a collection of paired data, which are the mosaicked noisy images $x$ and the denoised clean “ground truth” counterparts $y$. However, it is often costly and tedious to collect a large amount of high quality real-life data. Furthermore, the collected $y$ is not perfect without artifacts or noise. We illustrate this in Figure 1. For demosaicking, many approaches (Syu, Chen, and Chuang 2018; Tan et al. 2017b) take the output from a camera pipeline as $y$, possibly introducing artifacts like zipper effect or color moire in regions with rich textures and sharp edges. For denoising, the...
“clean” images are often collected by either setting a low-
ISO (Plotz and Roth 2017; Anaya and Barbu 2018) or aver-
ergating a set of repeated shots of the same scene (Abdel-
hamed, Lin, and Brown 2018), which still contain notice-
able noise. Moreover, such denoising data collecting process
usually assumes the captured objects to be perfectly still, or
requires a precise spatial alignment and intensity calibration
among a burst of images. Potential failure cases would intro-
duce additional error into the collected dataset. Therefore,
all these in-the-wild issues means that the “ground truth” $y$
deviates from the needed authentic $z$, limiting the perfor-
ance of DM&DN model.

To account for the fact that the collected ground truth $y$
is not a perfect reflection of $z$, we propose Wild-JDD, a
novel joint demosaicking and denoising learning framework
to enable training under ground truth uncertainty. In Wild-
JDD, we first formulate a two-stage data degradation pro-
cess, where a conjugate prior distribution is imposed upon
a base Gaussian distribution. Then, we derive an ELBO loss
from a variational perspective. In this way, the optimization
process is aware of the target uncertainty and prevents the
trained neural network from over-fitting to those randomness
errors. Beyond that, when the testing image falls outside of
the training range, we further enhance the performance by
regarding the input as a weakly informative prior.

Our main contributions are summarized as follows:

• We identify in existing DM&DN datasets the ground truth
uncertainty issues, manifesting themselves as various ar-
tifacts in the wild, such as zipper effect, color moire and
residual noise.

• We introduce a novel learning framework for joint demo-
saicking and denoising in the wild (Wild-JDD), where a
two-stage data degradation and an ELBO loss are formu-
lated for optimization. We also propose a simple but ef-
factive fine-tuning strategy for out-of-distribution input.

• Instead of simply generating a demosaicked clean image,
networks instantiated from our framework are capable of
estimating all the parameters involved in data degradation
and reconstruction, which provides better interpretability
of the optimization process.

• We conduct extensive experiments on both synthetic
and realistic datasets. Quantitative and qualitative com-
parisons show that Wild-JDD substantially outperforms
state-of-the-art works.

Related Work

In this section, we review the most relevant DM&DN works
from sequential processing to joint optimization, and from
supervised learning to self-supervised fine-tuning.

Traditionally, demosaicking and denoising are performed
sequentially in arbitrary order (Zhang et al. 2011; Akiyama,
Tanaka, and Okutomi 2015; Zhang et al. 2014). How-
ever, demosaicking first would correlate the noise spa-
tially, break the commonly used independent identically dis-
tributed (i.i.d.) assumption imposed on the noise modeling,
and increase the difficulty of the following denoising
step (Nam et al. 2016; Zhang et al. 2009). Another issue
arises if denoising is performed first, ending up with an
over-smoothed result (Jin, Facciolo, and Morel 2020). To
address the above problems, recent studies jointly consider
DM&DN for better performance. Khashabi et al. proposed
one of the very first joint approaches through a learned non-
parametric random field, and published an MSR dataset for
evaluation. Heide et al. embedded a non-local natural prior
into a global primal-dual optimization. Klatzer et al. for-
mulated a sequential energy minimization framework. How-
ever, these heuristics-based methods were outperformed by
the deep-learning-based approaches. Gharbi et al. trained
a neural network on millions of images to achieve better
results and shorter running time. After that, new approaches
were developed to extend the CNN’s capability in the field:
first with more effective network blocks (Huang et al. 2018;
Tan et al. 2018), or a cascade of refinement frameworks
(Kokkinos and Lefkimmiatis 2019, 2018), then by relying
on a variational deep image prior (Park et al. 2020), and ex-
ploring density map guidance to adaptively recover regions
with different frequencies (Liu et al. 2020). These learning-
based methods have achieved state-of-art performances, but
the assumption that the clean color image is perfect remains
in doubt.

Data-driven approaches normally perform well if the test-
ing image shares a similar distribution with the training data
(Mohseni et al. 2020; Ehret et al. 2019). However, in prac-
tice, the noise type remains diversified and may fall outside
the training range. For this reason, Lehtinen et al. introduced
a pioneering “noise2noise” training strategy using pairs of
noisy images. A similar “mosaic2mosaic” framework im-
proved a joint DM&DN network by fine-tuning bursts of
raw images (Ehret et al. 2019). Nonetheless, performance
of these methods can be limited by inadequate shots of the
same scene. Batson and Royer; Krull, Buchholz, and Jug
 tackled the problem by using only one realization for each
image at the price of a performance drop. These insightful
methods enable adaptive fine-tuning, but the quality or un-
certainty level of the pseudo ground truth has not been taken
into account.

Our method differs with previous works in that we do not
ideally assume the collected “ground truth” as perfect data.
Instead, we acknowledge the presence of various artifacts
and consider the case of training a joint DM&DN network
under the ground truth uncertainty.

Wild-JDD Methodology

We start with a dataset $D = \{x^{(i)}, y^{(i)}\}_{i=1}^{M}$ consisting
of $M$ pairs of images. Our goal is to learn a function that
maps $x^{(i)}$ to its corresponding authentic ground truth $z^{(i)}$
(note that $y^{(i)}$ is an approximation of $z^{(i)}$). This goal is
approached by two steps: (1) we formulate a two-stage data
degradation to link up all the parameters involved; (2) we
derive an expression of data likelihood for optimization.
Moreover, a fine-tuning strategy is applied to deal with
out-of-distribution inputs. To keep dimensional consistency
across different variables, in the following, without other-
wise stated, $x^{(i)}$ is first bilinearly interpolated to a color im-
age $\tilde{x}^{(i)}$ $\in \mathbb{R}^{3N}$, e.g. interpolating a missing green channel
by averaging its four green neighbors. The superscript \((i)\) is ignored in following subsections for simplicity.

**Two-stage Data Degradation**

Conventionally, the pixel/channel-wise expression for data degradation is modelled with an additive Gaussian noise (Zhou et al. 2019; Jia et al. 2019) as

\[
\tilde{x}_j | y_j, \sigma_j^2 \sim \mathcal{N}(y_j, \sigma_j^2),
\]

where \( j = 1, 2, \ldots, 3N \) specifies a dimension within an image. However, \( y_j \) is just a point estimator of the unknown authentic ground truth \( z_j \), and training with \( y_j \) only achieves a suboptimal performance as variance of this estimator is not considered. As such, we adopt a brand-new degradation model as

\[
\tilde{x}_j | z_j, \sigma_j^2 \sim \mathcal{N}(z_j, \sigma_j^2),
\]

and seek to first parameterize the unknown authentic ground truth \( z_j \). Suppose \( z_j \) follows a normal distribution with mean \( y_j \) and variance \( \sigma_j^2/\lambda \):

\[
z_j | y_j, \sigma_j^2, \lambda \sim \mathcal{N}(y_j, \sigma_j^2/\lambda),
\]

where \( \sigma_j^2 \) has an inverse gamma distribution parameterized by \( \alpha, \beta_j \):

\[
\sigma_j^2 | \alpha, \beta_j \sim \Gamma^{-1}(\alpha, \beta_j).
\]

Then \((z_j, \sigma_j^2)\) can be jointly denoted as a normal-inverse-gamma distribution:

\[
(z_j, \sigma_j^2) | y_j, \lambda, \alpha, \beta_j \sim \mathcal{N}-\Gamma^{-1}(y_j, \lambda, \alpha, \beta_j),
\]

which serves as a conjugate prior distribution over the base distribution in Equation (3). The collected ground truth \( y_j \)

\[
is an estimate of \( z_j \), while parameter \( \lambda \) reflects the quality or uncertainty level of this estimation. For \( \alpha \) and \( \beta_j \), they can be interpreted in a way that the variance \( \sigma_j^2 \) is estimated from \( 2\sigma \) observations with a sum of sample squared deviations \( 2\beta_j \). Similar to estimating the noise level by applying a Gaussian filter to the variance map (Yue et al. 2019), we parameterize \( \alpha \) and \( \beta_j \) as

\[
\alpha = \frac{w^2}{2},
\]

\[
\beta_j = \frac{w^2}{2} B(\{(x_j+t - y_{j+l})^2\}_{l=-[w/2]}^{[w/2]}),
\]

where \( B \) denotes a bilateral filtering operation on a variance map patch centered at pixel \( j \) with odd window size \( w \). Note that \( \lambda \) and \( \alpha \) here are invariant to the dimension \( j \). According to Equation (4), the variance of \( z_j \) or the uncertainty of \( y_j \) is proportional to the noise level \( \sigma_j^2 \). This can be reasonable, e.g., when collecting the clean ground truth by averaging multiple shots of the same scene, the noisier region should have lower quality and higher uncertainty (see Figure 2).

Therefore, based on equations (3) and (6), the data degradation process comprises two stages (see Figure 3): (1) \( z_j \) and \( \sigma_j^2 \) take values by sampling from the prior distribution \( p(z_j, \sigma_j^2 | y_j, \lambda, \alpha, \beta_j) \); (2) then a degraded sample \( \tilde{x}_j \) is generated from the conditional distribution \( p(\tilde{x}_j | z_j, \sigma_j^2) \).

**Maximizing ELBO**

For the purpose of predicting latent variables \( z, \sigma^2 \) given observed data point \( \tilde{x} \), or approximating the intractable posterior \( p(z, \sigma^2 | \tilde{x}) \), an encoder \( q_w(z, \sigma^2 | \tilde{x}) \) is introduced with learnable weights \( w \), as in the work (Kingma and Welling 2013). Notably, due to our use of a conjugate prior in Equation (6), this encoder \( q_w(z, \sigma^2 | \tilde{x}) \) is in the same probability distribution family as the prior, i.e.,

\[
q_w(z, \sigma^2 | \tilde{x}) = \prod_{i=1}^{3N} N^{-1}(\tilde{y}_j, \hat{\lambda}_j, \hat{\alpha}_j, \hat{\beta}_j),
\]

where \((\tilde{y}_j, \hat{\lambda}_j, \hat{\alpha}_j, \hat{\beta}_j)_{j=1}^{3N} \) is the output of a neural network. Note that here \( \hat{\lambda}_j \) is dimension-wise output learning the
same target $\lambda$ for different $j$. The same is applicable to $\hat{\alpha}_j$ and $\alpha$.

We next describe how we actually train such a neural network using a maximum likelihood estimation approach. As described in the work (Kingma and Welling 2013), the marginal log-likelihood can be calculated as

$$
\log p(\tilde{x}) = D_{KL}(q_w(z, \sigma^2|\tilde{x})||p(z, \sigma^2|\tilde{x})) + \mathcal{L}(w; \tilde{x}),
$$

where the KL divergence term is a non-negative value. Therefore, maximizing the marginal log-likelihood $\log p(\tilde{x})$ is converted to maximizing the second term, called the evidence lower bound (ELBO), and decomposed as

$$
\mathcal{L}(w; \tilde{x}) = -D_{KL}(q_w(z, \sigma^2|\tilde{x})||p(z, \sigma^2)) + \mathbb{E}_{q_w(z, \sigma^2|\tilde{x})} \left[ \log p(\tilde{x}|z, \sigma^2) \right].
$$

This ELBO loss is maximized when: (1) the divergence term encourages the distribution returned by the encoder network close to the prior; (2) the expectation term guides the network predicting parameters with a high encoder network after seeing the corrupted image. In the work (Kingma and Welling 2013), the sampling process requires a reparameterization trick for gradient back-propagation. However, such kinds of tricks are unnecessary here because a closed form expression for $\mathcal{L}(w; \tilde{x})$ can be derived analytically as follows:

$$
D_{KL}(q_w(z, \sigma^2|\tilde{x})||p(z, \sigma^2)) = \sum_{j=1}^{N} \left\{ \frac{\lambda \hat{\alpha}_j}{2 \beta_j} (y_j - \hat{y}_j)^2 + \frac{\lambda}{2 \lambda_j} - 2 \log \frac{\lambda}{\lambda_j} + \alpha_j \log \frac{\hat{\beta}_j}{\beta_j} - \frac{1}{2} \log \Gamma(\alpha_j) + \frac{\hat{\alpha}_j}{\beta_j} \right\},
$$

$$
\mathbb{E}_{q(z, \sigma^2|\tilde{x})} \left[ \log p(\tilde{x}|z, \sigma^2) \right] = \sum_{j=1}^{N} \left\{ -\frac{\log 2\pi}{2} - \frac{\hat{\beta}_j}{2} - \frac{\hat{\alpha}_j}{2 \lambda_j^2(\hat{\alpha}_j - 1)} + \frac{\hat{\beta}_j (\tilde{x}_j - \hat{y}_j)^2}{2 \beta_j} \right\},
$$

where $\Gamma(\cdot), \psi(\cdot)$ denote the Gamma and Di-gamma function respectively (detailed derivations are provided in the supplementary materials). Looking deeper into the term $\frac{\lambda \hat{\alpha}_j}{2 \beta_j} (y_j - \hat{y}_j)^2$ in Equation (11), we can notice that if the parameter $\lambda_j$ is set to be large enough, the ELBO loss would degenerate to a mean squared error (MSE). When using an MSE loss, too much attention would be put into the restoration term $(y_j - \hat{y}_j)^2$, leaving the existence of ground truth uncertainty omitted and making the model biased to the training data. Therefore, from a variational point of view, WildJDD provides a sound interpretability of the reason why the restoration term and the rest regularization terms should co-exist in the training process.

After formulating the ELBO loss for each single image, the overall optimization objective is obtained by computing across the entire dataset:

$$
\min_{w} \sum_{i=1}^{M} -\log p(\tilde{x}^{(i)}); \log p(\tilde{x}^{(i)}) \approx \mathcal{L}(w; \tilde{x}^{(i)}). \quad (13)
$$

At test time, the desired demosaicked clean image is obtained by taking the expectation of $z$, i.e. $\mathbb{E}[z] = \hat{y}$, while the noise map is parameterized as $\mathbb{E}[\sigma^2] = \frac{\hat{\beta}}{\hat{\alpha} - 1}$, according to the definition of N-$\Gamma^{-1}$ distribution.

**Corrupted Input as a Weakly Informative Prior**

Data-driven approaches are generally promising when a test image shares similar characteristics with the training set. However, their performances are limited when the input is considerably different, e.g. having a different type of noise. Inspired by the “noise2noise” (Lehtinen et al. 2018) and “mosaic2mosaic” (Ehret et al. 2019) algorithms, we further improve our trained model by taking the corrupted input as a weakly informative prior, i.e. replacing $y_j$ by $\tilde{x}_j$ during fine-tuning and using a smaller $\lambda$ to indicate the increased uncertainty. However, it comes with an underlying problem: the network may merely learn an identity mapping function, i.e. predicting $\tilde{x}_j$ given $\tilde{x}_j$.

It can be observed that in smooth regions, pixels share a strong spatial correlation. Therefore, we tackle the above issue by an alternative scheme, replacing $y_j$ by a random pixel $\tilde{x}_{j+t}$ in a small patch $\{\tilde{x}_{j+t} | t = [-p^2/2, p^2/2]\}$ centered at pixel $\tilde{x}_j$, where $p$ denotes the patch size. It is possible that $\tilde{x}_{j+t}$ differs a lot to $\tilde{x}_j$ in texture-rich regions, requiring another step to exclude the outlying $\tilde{x}_{j+t}$. We achieve this by applying a simple filter: if the value $\tilde{x}_{j+t}$ falls outside the confidence interval $(\tilde{x}_j - 2 \sigma_j, \tilde{x}_j + 2 \sigma_j)$, this informative prior $\tilde{x}_{j+t}$ is
In previous section, our network is trained to learn a mapping function
\[ f_t = \sum_{j=1}^{3N} \mathbb{I}(\tilde{x}_{j+t} \in (\tilde{x}_j - 2\sigma_j, \tilde{x}_j + 2\sigma_j))L_j(w; \tilde{x}), \quad (14) \]
where \( \mathbb{I}(\cdot) \) denotes an indicator function, and \( L_j(w; \tilde{x}) \) can be computed with \( j \)-indexed components in Equation (10), (11), (12) after replacing \( y_j \) by \( \tilde{x}_{j+t} \).

Illustrative Experimental Results

To show the effectiveness of our framework, we conduct extensive experiments with both synthetic datasets and realistic raw data. We focus on the Bayer pattern, which has been the dominating choice among various CFA patterns.

Network Architecture

In previous section, \( q_w(z, \sigma^2|x) \) represents the network taking \( \tilde{x} \) as the input. However, the original input is actually \( x \), and the bilinear interpolation process from \( x \) to \( \tilde{x} \) can be considered as part of the job done by the network. Therefore, our network is trained to learn a mapping function \( q_w(z, \sigma^2|x) \). We use a light-weight network architecture as shown in Figure 4. The GRDB building module refers to a grouped Residual Dense Block (Zhang et al. 2018), consisting of dense connected layers and a local feature fusion. A downscaling layer is positioned at the first layer to rearrange a mosaicked input to four quarter-resolution color maps. This rearrangement helps to save memory and speed up the training. Each of the first three convolution layers has 64 filters with \( 3 \times 3 \) kernel size. After that, an upscaling layer is used to unpack the features back to full-resolution. The last convolution layer produces 12 feature maps with \( 3 \times 3 \) kernel size. These 12 feature maps correspond to four parameters \( \tilde{y}, \tilde{\lambda}, \alpha, \beta \), each of which has 3 maps. The network is implemented using PyTorch framework.

Experiments on Synthetic Datasets

We first compare our method with previous works on the synthetic datasets in sRGB space, following the convention without an inverse ISP processing (Liu et al. 2020; Kläetz et al. 2016). In this experiment, 800 high-resolution images from DIV2K (Timofte et al. 2017) and 2650 high-resolution images from Flickr2K (Lim et al. 2017) are used for training. These images are randomly cropped into \( 1022 \times 120 \) patches with a batch size of 128. After the augmentation by flipping and rotation, the noisy mosaicked inputs are generated by applying the Bayer pattern sampling and adding random Gaussian noise in the range of \([0, 20]\). Unlike most denoising works assuming the i.i.d. noise, which deviates from the practical application, we adopt non-i.i.d Gaussian modeling with spatially variant noise levels, following the work (Yue
et al. 2019). During network training, we empirically set the parameter $\lambda$ as $2e3$ and $\alpha$ as $180.5$ (window size $w$ as $19$). The Adam optimizer (Kingma and Ba 2014) is used. The learning rate is initialized as $5e-4$, reduced by a factor of $0.8$ when the training meets a plateau in PSNR, with a minimum value of $1e-4$. The whole training process takes around 5 days on a single RTX 2080Ti GPU.

For testing, five widely-used benchmark datasets are used, including: Kodak\(^1\), McMaster (Zhang et al. 2011), WED-CMD (Tan et al. 2017b), MIT moire (Gharbi et al. 2016) and Urban100 (Huang, Singh, and Ahuja 2015). These datasets are collected from various devices under different scenarios. Note that the ground truth in these datasets are not perfect either. However, if one method can consistently outperform others across various datasets, its effectiveness can still be validated and approved. For comparison, six existing state-of-the-art works on joint DM&DN task are adopted, including: FlexISP (Heide et al. 2014), SEM (Klatzer et al. 2016), ADMM (Tan et al. 2017a), DeepJoint (Gharbi et al. 2016), Kokkinos (Kokkinos and Lefkimmiatis 2018) and SGNet (Liu et al. 2020). We run their source code for evaluation or directly cite their reported performance if the code is not available. The results are reported in both PSNR and SSIM matrices listed in Table 1.

Overall, our method outperforms all other works quantitatively, though our method is trained for non-i.i.d. noise cases. For both DeepJoint and SGNet, they assume an accurate noise map as known input, which is not reasonable in practice. In contrast, our method is able to perform a truly-blind reconstruction without such a known noise map. To further improve the performance, we adopt a self-ensemble strategy by applying the flipping and rotation on the input to generate 8 augmented inputs. After being processed by the network, 8 outputs are obtained and transformed back to the original geometry, followed by an averaging to get a unified final output. Note that the augmentation on input would break its Bayer pattern, e.g. from RGGB to GRBG. Therefore, a Bayer Preserving Unification is utilized by padding and cropping the image borders (Liu et al. 2019). We denote our method with self-ensemble as Wild-JDD*. Qualitative comparison is also provided in Figure 5. Our reconstructions remove the noise and preserve details without introducing noticeable artifacts, while other works tend to produce color moire in high-frequency regions. Although Kokkinos (Kokkinos and Lefkimmiatis 2018) also has good immunity to those artifacts, the produced images are over-smoothed due to their iterative processing properties.

### Experiments on Realistic Raw Data

In the previous experiment, we trained and evaluated on sRGB data to enable more comparison with other works. However, Khashabi et al. suggested that the evaluation should also be conducted on the raw data and thus proposed a realistic MSR 16-bits benchmark dataset. We retrain our network on their Linear Bayer Panasonic set with 200 images, in the same parameter setting to previous experiment. Table 2 reports our overall better performance in both linear and sRGB space.

<table>
<thead>
<tr>
<th>Method</th>
<th>linear PSNR</th>
<th>linear SSIM</th>
<th>sRGB PSNR</th>
<th>sRGB SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>JMCDM</td>
<td>37.44</td>
<td>0.971</td>
<td>31.35</td>
<td>0.942</td>
</tr>
<tr>
<td>RTF</td>
<td>37.77</td>
<td>0.976</td>
<td>31.77</td>
<td>0.951</td>
</tr>
<tr>
<td>FlexISP</td>
<td>38.28</td>
<td>0.974</td>
<td>31.76</td>
<td>0.941</td>
</tr>
<tr>
<td>SEM</td>
<td>38.93</td>
<td>0.980</td>
<td>32.93</td>
<td>0.960</td>
</tr>
<tr>
<td>DeepJoint</td>
<td>38.61</td>
<td>0.963</td>
<td>32.58</td>
<td>0.913</td>
</tr>
<tr>
<td>Kokkinos</td>
<td>39.29</td>
<td>0.975</td>
<td>33.37</td>
<td>0.930</td>
</tr>
<tr>
<td>MMNet20</td>
<td>40.07</td>
<td>0.979</td>
<td>34.24</td>
<td>0.942</td>
</tr>
<tr>
<td>DMCNN-VD</td>
<td>38.33</td>
<td>0.968</td>
<td>32.00</td>
<td>0.920</td>
</tr>
<tr>
<td>DMCNN-VD-Tr</td>
<td>40.07</td>
<td>0.981</td>
<td>34.08</td>
<td>0.957</td>
</tr>
<tr>
<td>Wild-JDD</td>
<td>40.16</td>
<td>0.980</td>
<td>34.34</td>
<td>0.945</td>
</tr>
<tr>
<td>Wild-JDD*</td>
<td><strong>40.36</strong></td>
<td><strong>0.981</strong></td>
<td><strong>34.59</strong></td>
<td><strong>0.947</strong></td>
</tr>
</tbody>
</table>

Table 2: Evaluation on realistic raw data. Our network is trained once using linear data and evaluated on both linear and sRGB space.

\(^1\)http://r0k.us/graphics/kodak
Figure 6: Increasing PSNR values when fine-tuning for different corrupted inputs. For each iteration, the updated PSNR values are obtained by averaging them across the McM dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Kodak PSNR</th>
<th>SSIM</th>
<th>McMaster PSNR</th>
<th>SSIM</th>
<th>WED-CDM PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>33.77</td>
<td>0.9124</td>
<td>33.45</td>
<td>0.9096</td>
<td>33.37</td>
<td>0.9233</td>
</tr>
<tr>
<td>ELBO</td>
<td>33.81</td>
<td>0.9127</td>
<td>33.53</td>
<td>0.9123</td>
<td>33.44</td>
<td>0.9244</td>
</tr>
</tbody>
</table>

Table 3: Comparison of ELBO against the MSE on synthetic datasets with noise level $\sigma = 10$.

Ablation Study

ELBO versus MSE  When setting the parameter $\lambda$ to be large enough, our ELBO loss would degrade to a commonly used MSE loss, which assumes the dataset ground truth to be a perfect target. The superiority of using ELBO against MSE loss is validated in Table 3, where a consistent PSNR improvement can be observed. This slight improvement comes from capturing the mild uncertainty embedded in the collected ground truth during training.

We also conduct an additional experiment to compare their optimization process. We take a single image from the Cartoon Set (Royer et al. 2020) as $z$. Then a Bayer pattern mosaicking and AWGN noise are applied to this image, followed by a bilinear interpolation to obtain a corrupted version $\hat{x}$. This corrupted image $\hat{x}$ is set to be the learning objective of a neural work. As described in the work Deep Image Prior (Ulyanov, Vedaldi, and Lempitsky 2018), a neural network tends to learn the clean signal faster than learning the random noise. Therefore, we can see that in Figure 7, the PSNR curve of MSE increases and then decreases. When ELBO comes into play, its PSNR curve fluctuates roughly around the curve of MSE. This fluctuation results from the interaction between ELBO’s restoration term and ELBO’s regularization terms. This is consistent to our expectation, that the neural network is aware of the uncertainty affiliated with the target $\hat{x}$ instead of treating $\hat{x}$ as the absolute learning target. Therefore, training with our ELBO loss can achieve a higher intermediate peak of the PSNR curve.

Effect of the mask  Using the mask during fine-tuning can effectively avoid edges getting blurred. As shown in Figure 8 (b), a mask map generated by our simple confidence interval scheme is able to outline those edges. In Figure 8 (c), fine-tuning with such a mask preserves the sharp edges more faithfully than fine-tuning without the mask.

Conclusion

We have presented the Wild-JDD, a novel learning framework for joint demosaicking and denoising tasks. We identify the ground truth uncertainty issues, formulate a two-stage data degradation process and derive an ELBO loss for optimization. We also propose a simple but effective fine-tuning strategy for out-of-distribution input. Comprehensive experiments demonstrate the effectiveness of our method. Wild-JDD not only outperforms state-of-the-art solutions in terms of both statistical and perceptual quality by a clear margin, but also provides good interpretability, where the restoration term and the rest regularization terms coexist to account for the learning target uncertainty. We hope that Wild-JDD will inspire more future research to study the effective training under the ground truth uncertainty in image-to-image translation tasks.
Acknowledgements
This work was supported, in part, by Hong Kong General Research Fund under grant number 16200120.

References


