Unsupervised Single Image Dehazing Using Dark Channel Prior Loss

Alona Golts
Department of Computer Science
Technion, Israel Institute of Technology
Technion City, Haifa 32000, Israel
salonaz@campus.technion.ac.il

Daniel Freedman
Google
Google Research, Haifa, Israel
danielfreedman@google.com

Michael Elad
Department of Computer Science
Technion, Israel Institute of Technology
Technion City, Haifa 32000, Israel
elad@cs.technion.ac.il

Abstract

Single image dehazing is a critical stage in many modern-day autonomous vision applications. Early prior-based methods often involved a time-consuming minimization of a hand-crafted energy function. Recent learning-based approaches utilize the representational power of deep neural networks (DNNs) to learn the underlying transformation between hazy and clear images. Due to inherent limitations in collecting matching clear and hazy images, these methods resort to training on synthetic data; constructed from indoor images and corresponding depth information. This may result in a possible domain shift when treating outdoor scenes. We propose a completely unsupervised method of training via minimization of the well-known, Dark Channel Prior (DCP) energy function. Instead of feeding the network with synthetic data, we solely use real-world outdoor images and tune the network's parameters by directly minimizing the DCP. Although our "Deep DCP" technique can be regarded as a fast approximator of DCP, it actually improves its results significantly. This suggests an additional regularization obtained via the network and learning process. Experiments show that our method performs on par with other large-scale, supervised methods.

1. Introduction

Haze is an atmospheric phenomenon where small particles, called aerosols, obstruct the clarity of an outdoor scene and lead to poor contrast and loss of detail. The existence of haze affects an image in two aspects. It attenuates the scene radiance with correspondence to an object's distance from the camera. Moreover, it introduces an additional ambient light component, called the *airlight*, which causes a "veiling effect" over the clear image. The formation of a hazy image is often described as a linear, per-pixel combination of the clear scene radiance and the airlight; the effect of each component is controlled by the transmission map. To recover the scene radiance image, one has to solve a system of $3N$ linear equations with $4N + 3$ unknowns (where N is the number of image pixels).

Figure 1: Results of our method, utilizing the DCP [8] for unsupervised training of a DNN, improving its results considerably and creating more natural looking images.

In order to handle the under-constrained haze creation model, many researchers suggested hand-crafted image priors, shedding additional light on the behaviour of hazy versus clean images [4, 29, 8, 30, 5, 33, 2, 17, 1]. These prior-based methods often formulate the problem of dehazing as an energy minimization task, where obtaining the solution

of each image is called "inference", requiring a non-trivial optimization scheme. With the increasing importance of image dehazing as an initial pre-processing stage in many computer-vision tasks (e.g., object detection, autonomous car navigation), large-scale, learning-based techniques have been deployed to solve it [15, 3, 22, 23, 31]. These methods, however, require thousands of input and output examples.

Since clean and hazy images of the exact same scene and lighting conditions are hard to obtain, learning-based methods commonly resort to synthetic dataset creation. Given a clean image and a corresponding depth map, one can calculate the transmission map and use the haze creation model to obtain hazy images with varying amounts of haze and airlight components. These pairs of hazy and clear images are later fed as inputs and labels in a supervised training of a DNN. Outdoor depth information, however, is incredibly imprecise. For instance, the depth information of the outdoor Make3D [24] and KITTI [6] datasets suffers from over 4 meters of average root-Mean-Square-Error (rMSE), while the rMSE of the indoor NYU2 [27] is only 0.5. Consequently, large-scale methods either use the more reliable indoor depth information [22, 15, 23, 31], or draw the depth map at random [33, 3]. Either of these practices creates a domain shift when addressing real-world outdoor images.

We propose to leverage the representational power of DNNs, but instead of feeding them with synthetic (often inaccurate) pairs of hazy and clean images, we train them in an unsupervised fashion using real-world hazy images only. We optimize the network's weights by minimizing an unsupervised loss function, which is essentially the Dark Channel Prior (DCP) [8] energy function. Our network can be regarded as a fast and simple, feed-forward approximator of the DCP. However, by stopping the optimization early, we get a significant boost in results over the classic DCP. This implies an added regularization, stemming from the network architecture and learning process. Our network, based on the Context Aggregation Network (CAN) architecture [32], is trained end-to-end from scratch, without relying on any external data apart from raw hazy images. It provides the predicted transmission maps as output, from which the dehazed image can be easily reconstructed.

Our method, that we call "Deep-DCP", is beneficial in several critical ways: (1) it is completely unsupervised, relying solely on real-world images; alleviating the domain shift caused by training on less-than-accurate indoor synthetic datasets, eliminating the need for them altogether; (2) as opposed to prior-based methods, it does not require a computationally-intensive optimization for each image, but rather learns the underlying transformation during training and requires only a forward-pass during test; (3) it harnesses the representational power of DNNs and uses them as additional regularization, on top of the already successful DCP. We perform a comprehensive quantitative evaluation of our method, and present state-of-the-art results on the *SOTS-outdoor* test set, in the recently released RESIDE dataset [16]. We show qualitative results on real-world images, demonstrating that the additional regularization provided by the network reduces common artifacts of prior-based methods, such as over-saturation and high-contrast.

To the best of our knowledge, our method is the first to employ unsupervised training of DNNs for the task of single image dehazing via energy-based loss optimization. The remainder of this paper is structured as follows: Section 2 provides a survey of previous prior-based and data-driven approaches for dehazing; Section 3 describes the DCP, its use as a loss function and our CAN-based architecture; Section 4 provides quantitative and qualitative experimental results; Section 5 includes dicussions and further analysis; finally, Section 6 concludes this work.

2. Previous Work

2.1. Prior-Based Approaches

Early attempts at image dehazing have incorporated several images of the same scene, taken at different bad weather conditions [19], or using different polarization filters [26]. Kopf *et al.* [12] later performed dehazing of outdoor images by utilizing existing geo-referenced terrain and urban models including depth, texture and GIS data.

In [29], Tan *et al.* unveil the haze from a single image by maximizing the local contrast of each patch in the image using a Markov Random Field (MRF) framework. In [4], Fattal *et al.* suggested utilizing the lack of correlation between the transmission and shading in a localized set of pixels, as a prior to resolve the ambiguity between the scene albedo and the airlight. Tarel *et al.* [30] provided a fast calculation of the "atmosperic veil", using a series of edge-preserving linear filter operations. In [20], Nishino *et al.* exploited the statistical independence between the scene albedo and depth, and factorized both quantities into an MRF-based energy function.

In [8], He *et al.* proposed the now widely used DCP, and demonstrated that in clear images the darkest pixel in an image patch is close to zero (this, however, does not hold in sky-regions). Using this and the assumption that the transmission map within a small image patch is constant, a coarse map can be easily derived. They further suggested a computationally costly soft matting operation for smoothing out the transmission and reconstructing the final dehazed image. Follow-up works have improved both the quality and efficiency of DCP. Specifically, in [17], the authors proposed a general boundary constraint for the transmission map, for which the DCP is a special case.

Several color-based priors have been suggested as well for boosting dehazing performance [5, 7, 2]. In [5], Fattal used the "color-lines" assumption, stating that pixels in

small image patches have a one-dimensional distribution in RGB-space [21]. There is an offset of these lines from the origin in hazy images, allowing to estimate the transmission map. Berman *et al.* proposed a global approach, called non-local dehazing (NLD) [2]. They observed that a haze-free image contains only several hundreds of distinct colors, clustered as points in RGB-space. In the presence of haze, these color clusters form a "haze-line", where the position of a certain pixel along the line corresponds to its initial radiance color and distance from the camera.

While prior-based methods reveal fine image details, they often suffer from increased saturation and contrast, unrealistic colors and difficulty in handling sky regions. This is due in part to assumptions not suited for all hazy image patches. In addition, each image requires a separate, non-trivial optimization and solution, which can be prohibitive for real-time applications.

2.2. Data-Driven Approaches

In [33], a Color Attenuation Prior (CAP) is suggested, mixing hand-crafted observations with a data-driven approach. CAP assumes that the image depth, the amount of haze and the difference between the brightness and saturation, are linearly correlated. To find the exact correlation, the authors opt for supervised regression between synthesized hazy patches and their corresponding depth maps. This results in fast inference at test time.

One of the first works to propose single image dehazing using CNNs is [22]. The method, called MSCNN, is trained by feeding a two-stage network with pairs of hazy images and corresponding transmission maps. In DehazeNet [3], Cai *et al.* create a novel CNN architecture (featuring max-out and BReLU layers), inspired by popular prior-based methods [8, 29, 33, 1]. AOD-Net [15] in turn, proposes a joint estimation of both the transmission map and the airlight via a unified representation. Using this representation, one can easily reconstruct the scene radiance directly in an end-to-end forward-pass computation. This helps reduce errors accumulated in the separate calculation of the two quantities.

In the recent Gated Fusion Network (GFN) [23], a dehazed image is produced as a fusion of the white balance, contrast enhanced and gamma corrected images (all derived from the hazy image). The network outputs three confidence maps which determine the effect of each component. To combat halo effects of a single scale encoder-decoder structure, a multi-scale architecture is used, where a coarse output is first produced, then added as input to a finer scale network. This method provides impressive results on RESIDE's *SOTS-indoor*, but quadruples the size of the input during training and test, making evaluation inefficient in terms of memory. Finally, in a recent work reported in [31], the authors utilize the pre-trained VGG [28] network as encoder and only train the symmetric decoder via a combination of MSE and perceptual loss.

While learning-based methods achieve impressive results, they are trained in a supervised way, relying on synthetic datasets. Some methods use more accurate *indoor* depth information to create labelled inputs [22, 15, 23, 31]. This practice, however, directs increasing research effort to optimizing indoor performance, while the predominant need for dehazing is actually outdoors. Other methods use real-world *outdoor* images, but compromise the accuracy of the depth information. For example, [33] draws each pixel in the depth map at random from a $(0, 1)$ uniform distribution, and [3] enforces an additional constraint of constant depth within 16×16 patches. These assumptions result in block and halo artifacts in the reconstructed image, and require additional post-processing.

3. Our Method

In the following we will describe our method for single image dehazing, including the driving force of our unsupervised loss function – the Dark Channel Prior [8] – its implementation as a loss function for training a CNN, and the architecture we choose for the task at hand.

3.1. Haze Model

The popular haze formation model in [18] is given as:

$$\begin{aligned} \mathbf{I}(\mathbf{x}) &= t(\mathbf{x})\mathbf{J}(\mathbf{x}) + (1 - t(\mathbf{x}))\mathbf{A}, \\ t(\mathbf{x}) &= e^{-\beta d(\mathbf{x})}. \end{aligned} \tag{1}$$

According to the above, the observed hazy image, $\mathbf{I}(\mathbf{x}) \in \mathbb{R}^{N\times 3}$, is a convex linear combination of the haze-free scene radiance, $\mathbf{J}(\mathbf{x})$, and the atmospheric light component, $\mathbf{A}$, called the *airlight*; usually represented as a constant 3-vector in RGB-space, $\mathbf{A} = \left(A^r, A^g, A^b\right)$. The transmission map coefficients, $t(\mathbf{x}) \in \mathbb{R}^N$ control the relative force of each component, in each pixel in the image, $\mathbf{x} \in \mathbb{R}^N$. The transmission is a function of the depth, $d(\mathbf{x})$, of the scene from the observer. Our goal in single-image-dehazing is to obtain the haze-free scene radiance, $\mathbf{J}(\mathbf{x})$. To do so, however, one needs to solve a set of $3N$ equations (only $\mathbf{I}(\mathbf{x})$ is given), with $4N+3$ unknowns ($\mathbf{J}(\mathbf{x}), t(\mathbf{x}), \mathbf{A}$). Thus, additional prior knowledge of the images in question is needed.

3.2. Dark Channel Prior

The dark channel prior is an image statistical property, indicating that in small patches of haze-free outdoor images, the darkest pixel across all color channels is very dark, and close to zero. The "dark channel" of the image is defined as

$$\mathbf{J}^{dark}(\mathbf{x}) = \min_{c\in\{r,g,b\}} \left(\min_{\mathbf{y}\in\Omega(\mathbf{x})} \left(\mathbf{J}^c(\mathbf{y})\right)\right), \tag{2}$$

where $\Omega(\mathbf{x})$ is a small patch, centered around $\mathbf{x}$. This observation is contributed by three factors which appear in

outdoor images: (1) shadows – induced by cars, buildings and trees; (2) colorful objects – where one color channel is dominant, and the others are close to zero, e.g., red flowers, green leaves, blue sea; and (3) naturally dark objects – such as tree trunks and rocks.

Assuming that $\mathbf{A}$ is known and the transmission within a small image patch, denoted as $\tilde{t}(\mathbf{x})$, is constant, one can apply a minimum operation across channels and pixels in the haze formation equation in (1) (effectively zeroing $\mathbf{J}^c(\mathbf{y})$) and get a prediction for the transmittance [8]:

$$\tilde{t}(\mathbf{x}) = 1 - \omega \cdot \min_c \left(\min_{\mathbf{y} \in \Omega(\mathbf{x})} \left(\frac{\mathbf{I}^c(\mathbf{y})}{A^c} \right) \right), \tag{3}$$

where $\omega = 0.95$ leaves a small amount of haze for natural-looking results. In sky regions although the dark channel does not always hold, it is assumed that $\mathbf{I}/\mathbf{A} \rightarrow 1$, thus $\tilde{t}(\mathbf{x}) \rightarrow 0$. The resulting coarse transmission map requires an additional step of refinement.

3.3. Soft Matting

The haze formation model in (1) is very similar to the composition model in image matting [14], where an output image is a convex linear combination of foreground and background images; controlled by the alpha matte, α. If one replaces the α-matte with the coarse transmission map, $\tilde{t}(\mathbf{x})$, the following energy function suggested in [14] can be used to acquire the refined map, $t(\mathbf{x})$:

$$E(\mathbf{t}, \tilde{\mathbf{t}}) = \mathbf{t}^T \mathbf{L} \mathbf{t} + \lambda (\mathbf{t} - \tilde{\mathbf{t}})^T (\mathbf{t} - \tilde{\mathbf{t}}), \tag{4}$$

where the first term promotes successful image matting, and the second, fidelity to the dark channel solution. The parameter λ, controlling the force between the two, is set to $\lambda = 10^{-4}$ [8]. The matrix $\mathbf{L}$ is a Laplacian-like matrix, dedicated to image matting and given by [14]:

$$\begin{aligned} \mathbf{L}_{ij} &= \sum_{n|(i,j)\in p_n} (\delta_{ij} - w_{ij}^n), \quad \forall i,j = 1...N \\ w_{ij}^n &= \tfrac{1}{|p_n|}\big(1 + (\mathbf{I}_i - \boldsymbol{\mu}_n)^T (\boldsymbol{\Sigma}_n + \tfrac{\varepsilon}{|p_n|}\mathbf{U}_3)^{-1} (\mathbf{I}_j - \boldsymbol{\mu}_n)\big), \end{aligned} \tag{5}$$

where i, j are two pixels within a small patch p_n around pixel n; $|p_n|$ is the size of the patch and equal to $3 \times 3 = 9$ as suggested in [14]; $\boldsymbol{\mu}_n \in \mathbb{R}^3$ and $\boldsymbol{\Sigma}_n \in \mathbb{R}^{3\times 3}$ are the mean and covariance of the patch; $\mathbf{U}_3$ is the identitly matrix; and ε is a smoothing parameter set to $\varepsilon = 10^{-6}$ [14].

3.4. Implementation as a Loss Function

We rewrite the energy function in equation (4) in a tensor-friendly format by using a known decomposition of Laplacian matrices via their weights, given in (5). Rephrasing the first term in (4) in terms of weights, we have that

$$E_1(\mathbf{t}, \tilde{\mathbf{t}}) = \mathbf{t}^T \mathbf{L} \mathbf{t} = \sum_{n=1}^{N} \sum_{i=1}^{9} \sum_{j=1}^{9} w_{ij}^n (t_i - t_j)^2, \tag{6}$$

where we sum over all overlapping patches around N pixels in the resulting transmission map, $\mathbf{t}$, as well as over all possible combinations of pixel pairs, i, j, in a given 3×3 patch. The maximum number of combinations is $(3^2) \cdot (3^2) = 81$. We can vectorize this term, along with the data fidelity term

$$E(\mathbf{t}, \tilde{\mathbf{t}}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{W} \odot (\mathbf{T}_I - \mathbf{T}_J)^2 + \lambda \sum_{n=1}^{N} (\mathbf{t} - \tilde{\mathbf{t}})^2, \tag{7}$$

where $\odot$ denotes elementwise multiplication; $k \in [1..81]$ indexes all possible pairs of pixels in a 3×3 patch, and $\mathbf{W} \in \mathbb{R}^{N \times 81}$ is the vectorized version of the weights. $\mathbf{T}_I, \mathbf{T}_J \in \mathbb{R}^{N \times 81}$ are repetitions of the output transmission map. The first representing the transmission patches (9 pixels in total) arranged in $I \rightarrow (1, .., 1, 2, ..., 2, ..., 9, ..., 9) \in \mathbb{R}^{81}$, and the second arranged in $J \rightarrow (1, 2, ..., 9, 1, 2, ..., 9, ..., 1, 2, ...9) \in \mathbb{R}^{81}$.

Above is the loss function with which we train our network, whose predicted transmission map is parametrized by $\mathbf{t}_\theta$. We tune the parameters, θ, by minimizing the loss function in (7) over the training set of hazy images, $\{\mathbf{I}_m\}_{m=1}^M$:

$$\theta^* = \arg\min_\theta \left[\frac{1}{M} \sum_{m=1}^{M} E(\mathbf{t}_\theta, \tilde{\mathbf{t}}(\mathbf{I}_m)) \right], \tag{8}$$

where M is the number of images. Note that we do not use the "labels", i.e., the clear images, at any point, only the original hazy ones. A schematic diagram of the inputs and outputs of our loss module is given in Figure. 3.

3.5. Computing the Scene Radiance

Once the network has finished training, the transmission map, $t_\theta(\mathbf{x})$, of a new hazy image can be obtained by a forward-pass operation. This is used to recover the scene radiance via the haze formation model in (1):

$$\mathbf{J}(\mathbf{x}) = \frac{\mathbf{I}(\mathbf{x}) - \mathbf{A}}{\max(t_\theta(\mathbf{x}), t_0)} + \mathbf{A}, \tag{9}$$

where t_0, which discourages division by numbers close to zero, is set to $t_0 = 0.1$ as suggested in [8]. In order to recover the missing airlight component, $\mathbf{A}$, we follow the method suggested in [8]: we first pick the 0.1% brightest pixels in the dark channel of the hazy image. Then, out of these locations we pick the brightest pixel in the hazy image, $\mathbf{I}$. That is the final chosen atmospheric light, $\mathbf{A}$.

3.6. Architecture

Our fully-convolutional, "Dilated Residual Network", shown in Figure 2, is inspired by the Context Aggregation Network (CAN) [32], which has shown impressive results in dense-output applications. Similarly to CAN, we keep the resolution of all layers intact and identical to that of the input and output. In order to get an accurate prediction we

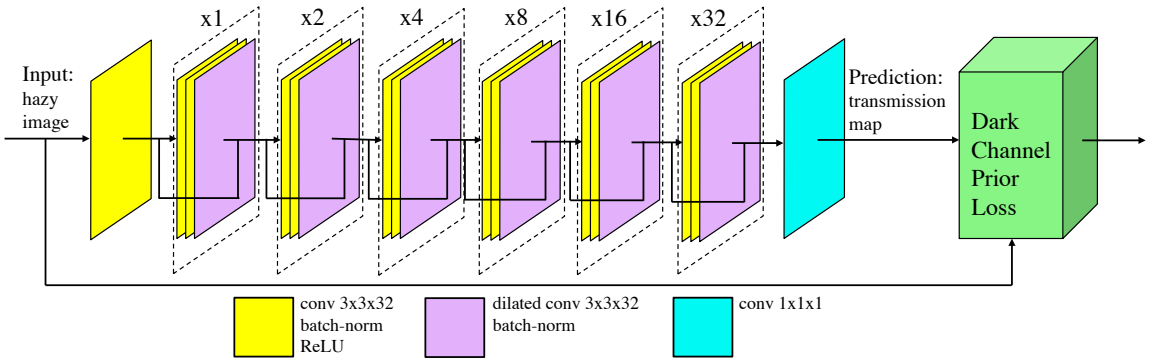

Figure 2: System architecture. Our fully-convolutional network receives real-world hazy images. Apart from the input and output layers, our network is a cascade of dilated residual blocks (dilation written above each block), which gradually increase the receptive field. The network's predicted transmission and the input image, are fed to the unsupervised, DCP loss.

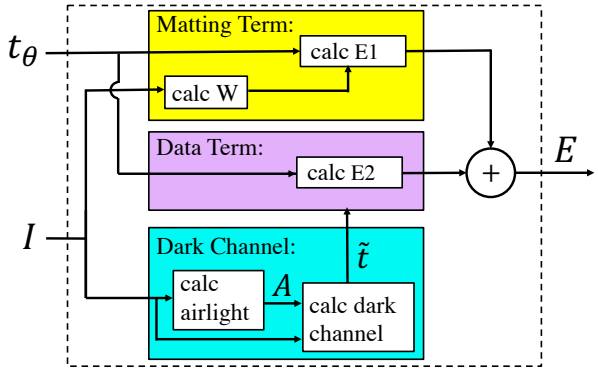

Figure 3: Our loss module, which receives the prediction of the network, $\mathbf{t}_\theta$, along with the hazy image, $\mathbf{I}$, and outputs the value of the DCP [8] energy loss.

avoid pooling and upsampling, and instead increase the receptive field via dilated convolutions with exponentially increasing dilation factors. Contrary to [32], between each dilated convolution we add another two regular convolution layers to create a richer nonlinear representation.

Our network is thus built as a cascade of 6 dilated residual blocks; each made up of two regular convolutions, followed by a single dilated convolution. The dilation factors increase by a power of two from one block to the next. The filter size and width of all convolution layers (apart from the output) is $3 \times 3 \times 32$. All regular convolutions are followed by batch normalization [10] and ReLU nonlinearity [13], and all dilated ones are followed by batch-norm only. The final layer is a linear transformation to the output dimension of the transmission map $1 \times 1 \times 1$. To improve gradient flow and propagate finer details to the output, we incorporate additional Resnet-style [9] skip connections between the input and output of each block. The skip connection is a simple addition of the input to the output of each block.

4. Experimental Results

4.1. Dataset

In order to train and evaluate the performance of our network, we use the recent large-scale RESIDE (REalistic Single Image DEhazing) dataset [16]. RESIDE's training set, called *"ITS"*, includes $13,990$ synthetic indoor images, created from the NYU2 [27] and Middlebury stereo datasets [25]. The test set includes both indoor and outdoor sections, called *"SOTS-indoor"* and *"SOTS-outdoor"*; each containing 500 synthetic images. A smaller test set of 20 outdoor images, called *"HSTS"*, is also suggested. HSTS has a mix of 10 synthetic images (where ground truth is known) and 10 real-world images. All synthetic hazy images are created by first collecting ground-truth clean images with their corresponding depth maps, and then applying the haze formation model with different configurations of the $\mathbf{A}, \beta$ parameters in (1). The beta version of RESIDE provides an additional collection of $4,322$ real-world images, mined from the web, called *"RTTS"*. Instead of using the synthetic indoor database of ITS (or its variations based on NYU2 and Middlebury), as in [23, 22, 15, 31], we train our network on the real-world images of RTTS. For the evaluation of PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity) criteria during training, we use a subset of 500 images, chosen randomly from ITS.

4.2. Implementation Details

To enrich the RTTS training set, we perform data augmentation. The first augmentation is simply resizing the original hazy images to size 128×128 using bilinear interpolation. The second, third and fourth augmentations are performed randomly. Each image can be flipped horizontally or kept as is; randomly cropped to 256×256 or 512×512, and rotated at $0, 45, 90$, or 135 degrees. If rotated, only the valid center of the image is taken. All aug-

	DCP [8]	BCCR [17]	NLD [2]	CAP [33]	MSCNN [22]	DehazeNet [3]	AOD-Net [15]	GFN [23]	Ours
HSTS	15.96/0.877	15.09/0.738	17.62/0.792	21.54/0.867	18.29/0.841	**24.49**/0.915	21.58/0.922	22.94/0.874	24.41/**0.934**
SOTS-outdoor	16.96/0.886	15.49/0.781	18.07/0.802	22.30/0.914	19.56/0.863	22.72/0.858	21.34/0.924	21.49/0.838	**24.07/0.933**
SOTS-indoor	19.79/0.848	16.88/0.791	17.29/0.749	19.05/0.836	17.11/0.805	21.14/0.847	19.38/0.849	**22.32/0.880**	19.25/0.832

Table 1: Quantitative PSNR/SSIM results of our approach (higher is better). For both SOTS-outdoor and HSTS we report the result of epoch 27, whereas in SOTS-indoor we report the result of epoch 30.

mented images are then resized to 128×128. The final number of training images is therefore: $4322 \times 4 = 17,288$.

The parameters of our loss function are taken exactly (no additional tuning) as suggested in [8, 14]: $\lambda = 10^{-4}$, $\omega = 0.95$, $t_0 = 0.1$, $\varepsilon = 10^{-6}$, DCP patch size: 15×15, and soft matting patch size: 3×3. We use the Adam optimizer [11] with batch size of 24; initial learning rate of $l_r = 3 \cdot 10^{-4}$; and exponential decay with factor 0.96 every 3 epochs. The network weights are initialized using random initialization with zero mean and variance of 0.1. Our method is implemented in TensorFlow on a GTX Titan-X Nvidia GPU. Training time to get the optimal solution (about 30 epochs) takes 8 hours, while evaluation of a 640×480 image takes 0.56 seconds. For outdoor results we stop the training at epoch 27, whereas for SOTS-indoor, we keep training until we reach 30 epochs. Our stopping criterion is explained further in section 5.2.

4.3. Quantitative Evaluation

We evaluate the performance of our method on the SOTS-indoor, SOTS-outdoor and HSTS test sets. These test sets are created synthetically, therefore featuring both the clean images and their hazy versions. We measure the quality of our solution in terms of the PSNR and SSIM metrics. We obtain the original code and compare our results to the following prior-based approaches: DCP [8] (using our own baseline implementation), BCCR [17] and NLD [2], and the following data-driven methods: CAP [33], MSCNN [22], DehazeNet [3], AOD-Net [15] and GFN [23]. We normalize the predicted solutions of all methods to $[0, 1]$ if that improves PSNR and SSIM.

The numeric results[1] are given in Table 1. We get the highest PSNR and SSIM scores among all other methods in the larger SOTS-outdoor, and the highest SSIM in the smaller HSTS. Our method, represented by a rich neural network and trained to accommodate numerous images, obtains better results compared to prior-based methods. Specifically, compared to DCP, our method strives to approximate the solution of the same energy function, but we stop it before reaching an absolute minimum, in order to get further regularization. This is particularly noticed in outdoor images, where DCP often over-saturates the sky.

With regard to data-driven approaches, our high score is attributed to the fact that we train on *real-world outdoor* images, whereas competing methods [22, 15, 23] concentrate on *synthetic indoor* images and suffer from a certain domain shift when addressing outdoor data. In addition, the synthetic hazy and clean pairs are created from coarse depth data, for which training creates a negative bias towards data-driven approaches. An example of an indoor training image in ITS is given in Figure 6. Notice the rough misplaced edges in the transmission map, which later translate to inaccurate hazy images. Indeed, our closest competitor in terms of outdoor results is DehazeNet [3]. Recall that this method is trained on a large variety of clean image patches of outdoor scenes, making it more robust compared to methods trained on ITS.

We include the results of our method on SOTS-indoor, where it performs favourably, but gets a lower score compared to other data-driven methods and even DCP. This is expected since we train on outdoor images, creating a trade-off between indoor and outdoor performance. As for DCP, it behaves more agreeably on indoor images which coincide better with the haze formation model and do not include objects located at infinity.

4.4. Qualitative Results

We present qualitative results on HSTS in Figure 4. In the top part of Figure 4, it can be seen that our method maintains the true colors of the original image, whereas DCP [8], BCCR [17] and NLD [2] tend to produce exaggerated sky regions. Our results are similar to those produced by CAP [33], however slightly closer to the true colors exhibited in the ground truth image. In the bottom half of Figure 4 we provide a comparison to deep-learning based methods. In most images we maintain the true contrast and colors, whereas MSCNN [22] and GFN [23] provide more contrast-enhanced images. At times, we slightly change the color of the sky, which is to be expected since our method is unsupervised and does not witness the clear images at any stage. In Figure 5, one can see a real-world' image comparison of our results with both prior-based and data-driven methods. We display the output of our network after 27 epochs (the optimal results for SOTS-outdoor) and after 30 epochs, where the produced images are more similar to DCP (see discussion on sec. 5.1). One can see that after 27 epochs we do not remove all of the haze, perhaps indicating that the outdoor images in RESIDE are less hazy than real-world hazy images. For 30 epochs, our result is more

[1] We get slightly different results than reported in [16].

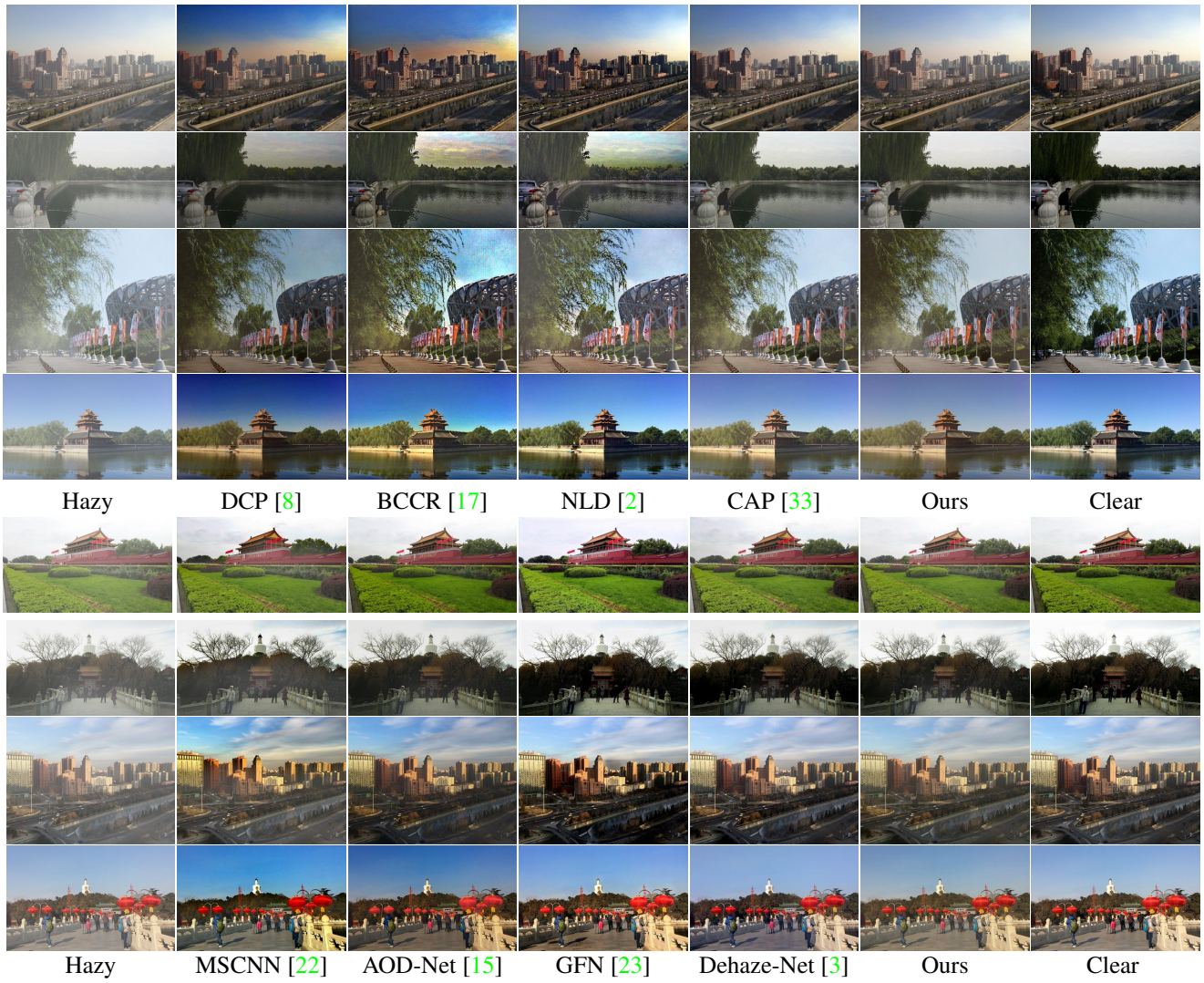

Figure 4: Qualitative results on RESIDE's HSTS. Upper half: comparison to prior-based methods; bottom half: comparison to deep-learning-based methods.

saturated, with higher contrast.

5. Discussion

5.1. Proximity to Dark Channel Prior

During training, our network strives to approximate the DCP energy function. Since it optimizes the loss for the entire corpus of images, it may output different results from DCP [8]. While DCP operates on one image at a time, our network learns a more "universal solution", suited for multiple images. In addition, as the epochs evolve and the loss value decreases, we reach closer and closer to DCP, as can be seen in the three rightmost columns in Figure 5. At earlier epochs the output images still contain a large amount of haze, whereas further on, most of the haze is lifted, but the colors appear more saturated, even non-realistic. We search for a middle-ground, where most of the haze is removed and one can see the details, but the colors and contrast remain realistic and physically valid. The benefit of stopping before reaching a deeper optimum of DCP is especially noticeable in sky regions, where DCP would output an exaggerated and amped-up version of the sky, whereas we produce a more natural color. In our case this "sweet-spot" is reached after 27 epochs over the training data. Nonetheless, we can keep training for a few more epochs to get more vivid results, which may be more pleasant to the human eye.

5.2. Unsupervised Training Regime

Although our training is completely unsupervised, we do need a stopping criterion since reaching the minimum of the

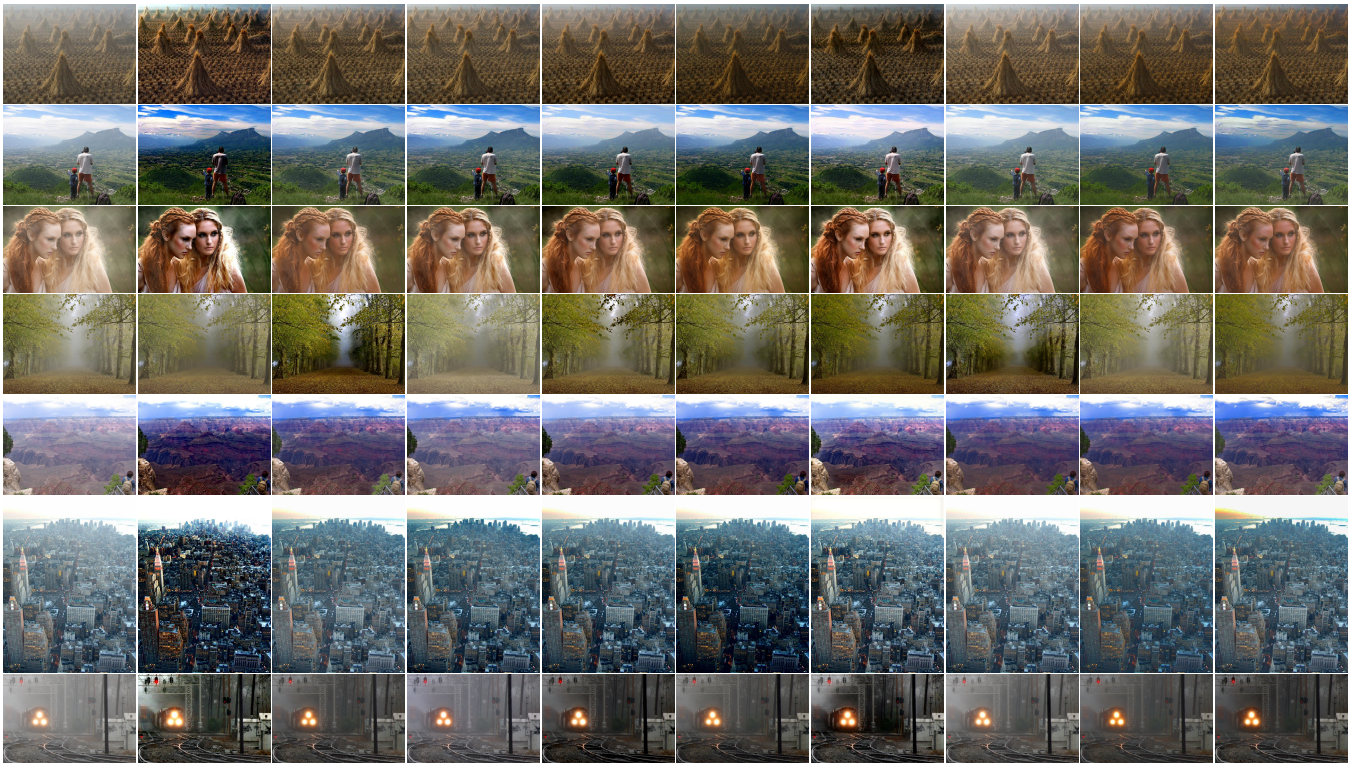

Figure 5: Qualitative results of single image dehazing on real-world images. The numbers in parenthesis are the number of epochs spent training our network.

Figure 6: Example of a synthetic image and its coarse transmission map from RESIDE's ITS training dataset [16].

loss function is not always beneficial in terms of the visual and quantitative results. To do so, we evaluate the averge loss value, PSNR and SSIM of a small supervised set of 500 images from ITS. A typical behaviour of the results is a decrease of the average loss; an increase in performance in PSNR and SSIM; reaching a maximum, and then, a decrease of these criteria. We save 10 epochs around the optimum, then choose the one that gave the best test performance on SOTS-outdoor. This is usually 1-3 epochs before the optimum results on ITS. The learning parameters are chosen using a similar technique.

6. Conclusions

We have presented a method of unsupervised training of deep neural networks for the purpose of single image dehazing. Our method relies on the well-known Dark Channel Prior (DCP) [8], and manages to improve it considerably. In addition to providing state-of-the-art performance in outdoor scenarios, our method also eliminates the need for synthetic training sets. While our focus here is DCP, we could have incorporated any other successful energy function, using it as our unsupervised loss. Our future research is focused on finding an even better *combination* of energy functions, or incorporating some amount of supervision to benefit from both worlds.

References

[1] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert. A fast semi-inverse approach to detect and remove the haze from a single image. In *ACCV*, 2010. 1, 3

[2] D. Berman, T. Tali, and S. Avidan. Non-local image dehazing. In *CVPR*, 2016. 1, 2, 3, 6, 7

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 25(11):5187–5198, 2016. 2, 3, 6, 7

[4] R. Fattal. Single image dehazing. *TOG*, 27(3):72, 2008. 1, 2

[5] R. Fattal. Dehazing using color-lines. *TOG*, 34(1):13, 2014. 1, 2

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[7] K. B. Gibson and T. Q. Nguyen. An analysis of single image defogging methods using a color ellipsoid framework. *EURASIP Journal on Image and Video Processing*, 2013(1):37, 2013. 2

[8] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2011. 1, 2, 3, 4, 5, 6, 7, 8

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[12] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski. Deep photo: Model-based photograph enhancement and viewing. *TOG*, 27(5), 2008. 2

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[14] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, 2006. 4, 6

[15] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017. 2, 3, 5, 6, 7

[16] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Reside: A benchmark for single image dehazing. *arXiv preprint arXiv:1712.04143*, 2017. 2, 5, 6, 8

[17] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *ICCV*, 2013. 1, 2, 6, 7

[18] W. K. Middleton. Vision through the atmosphere. In *Geophysik II/Geophysics II*, pages 254–287. Springer, 1957. 3

[19] S. G. Narasimhan and S. K. Nayar. Chromatic framework for vision in bad weather. In *CVPR*, 2000. 2

[20] K. Nishino, L. Kratz, and S. Lombardi. Bayesian defogging. *IJCV*, 98(3):263–278, 2012. 2

[21] I. Omer and M. Werman. Color lines: Image specific color representation. In *CVPR*, 2004. 3

[22] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016. 2, 3, 5, 6, 7

[23] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang. Gated fusion network for single image dehazing. *CVPR*, 2018. 2, 3, 5, 6, 7

[24] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 31(5):824–840, 2009. 2

[25] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003. 5

[26] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar. Polarization-based vision through haze. *Applied optics*, 42(3):511–525, 2003. 2

[27] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[29] R. T. Tan. Visibility in bad weather from a single image. In *CVPR*, 2008. 1, 2, 3

[30] J.-P. Tarel and N. Hautiere. Fast visibility restoration from a single color or gray level image. In *ICCV*, 2009. 1, 2

[31] Z. Xu, X. Yang, X. Li, X. Sun, and P. Harbin. Strong baseline for single image dehazing with deep features and instance normalization. *BMVC*, 2018. 2, 3, 5

[32] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2, 4, 5

[33] Q. Zhu, J. Mai, L. Shao, et al. A fast single image haze removal algorithm using color attenuation prior. *TIP*, 24(11):3522–3533, 2015. 1, 2, 3, 6, 7