AOD-Net: All-in-One Dehazing Network

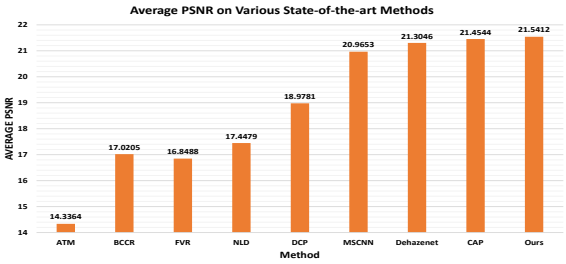
Boyi Li[1]*, Xiulian Peng[2], Zhangyang Wang[3], Jizheng Xu[2], Dan Feng[1]
[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[2]Microsoft Research, Beijing, China
[3]Department of Computer Science and Engineering, Texas A&M University
boyilics@gmail.com, xipe@microsoft.com, atlaswang@tamu.edu, jzxu@microsoft.com, dfeng@hust.edu.cn

Abstract

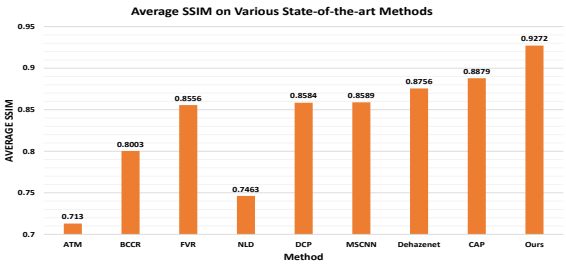
This paper proposes an image dehazing model built with a convolutional neural network (CNN), called All-in-One Dehazing Network (AOD-Net). It is designed based on a re-formulated atmospheric scattering model. Instead of estimating the transmission matrix and the atmospheric light separately as most previous models did, AOD-Net directly generates the clean image through a light-weight CNN. Such a novel end-to-end design makes it easy to embed AOD-Net into other deep models, e.g., Faster R-CNN, for improving high-level tasks on hazy images. Experimental results on both synthesized and natural hazy image datasets demonstrate our superior performance than the state-of-the-art in terms of PSNR, SSIM and the subjective visual quality. Furthermore, when concatenating AOD-Net with Faster R-CNN, we witness a large improvement of the object detection performance on hazy images.

1. Introduction

The existence of haze dramatically degrades the visibility of outdoor images captured in the inclement weather and affects many high-level computer vision tasks such as detection and recognition. These all make single-image haze removal a highly desirable technique. Despite the challenge of estimating many physical parameters from a single image, many recent works have made significant progress towards this goal [1, 3, 17]. Apart from estimating a global *atmospheric light* magnitude, the key to achieve haze removal is to recover a *transmission matrix*, towards which various statistical assumptions [8] and sophisticated models [3, 17] have been adopted. However, the estimation is not always accurate, and some common pre-processing such as guild-filtering or softmatting will further distort the hazy image generation process [8], causing sub-optimal restoration performance. Moreover, the non-joint estimation of two critical parameters, transmission matrix and atmospheric light, may further amplify the error when applied together.



(a) Comparison on PSNR

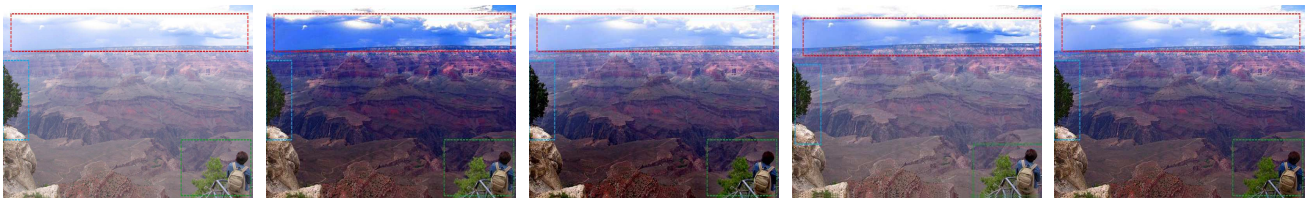


(b) Comparison on SSIM

Figure 1. The PSNR and SSIM comparisons on dehazing 800 synthetic images from Middlebury stereo database. The results certify that AOD-Net presents more faithful restorations of clean images.

In this paper, we propose an efficient end-to-end dehazing convolutional neural network (CNN) model, called *All-in-One Dehazing Network* **(AOD-Net)**. While some previous haze removal models discussed the "end-to-end" concept [3], we argue the major novelty of AOD-Net as the first to optimize *the end-to-end pipeline from hazy images to clean images*, rather than *an intermediate parameter estimation step*. AOD-Net is designed based on a re-formulated atmospheric scattering model. It is trained on synthesized hazy images, and tested on both synthetic and real natural images. Experiments demonstrate the superiority of AOD-Net over several state-of-the-art methods, in terms of not

*The work was done at Microsoft Research Asia.



(a) Inputs (b) DCP [8] (c) DehazeNet [3] (d) MSCNN [17] (e) AOD-Net

Figure 2. Visual quality comparison between AOD-Net and several state-of-the-art methods on a natural hazy image. Please amplify figures to view the detail differences in bounded regions.

only PSNR and SSIM (see Figure 1), but also visual quality (see Figure 2). As a lightweight model, AOD-Net has achieved a fast processing speed, costing as low as 0.026 second to process one 480×640 image with a single GPU. Furthermore, we are the first to examine how a haze removal model could be utilized to assist the subsequent high-level vision task. Benefiting from the end-to-end formulation, AOD-Net is easily embedded with Faster R-CNN [16] and improves the object detection performance on hazy images with a large margin.

2. Related Work

Physical Model: The *atmospheric scattering model* has been the classical description for the hazy image generation process [11, 13, 14].

$$I(x) = J(x)\,t(x) + A\,(1 - t(x)), \tag{1}$$

where $I(x)$ is observed hazy image, $J(x)$ is the scene radiance ("clean image") to be recovered. There are two critical parameters: A denotes the global atmospheric light, and $t(x)$ is the transmission matrix defined as:

$$t(x) = e^{-\beta d(x)}, \tag{2}$$

where β is the scattering coefficient of the atmosphere, and $d(x)$ is the distance between the object and the camera.

Traditional Methods: [23] coped with haze removal by maximizing the local contrast. [6] proposed a physically-grounded method by estimating the albedo of the scene. [8, 24] discovered the effective dark channel prior (DCP) to more reliably calculate the transmission matrix. [12] further enforced the boundary constraint and contextual regularization for sharper restored images. An accelerated method for the automatic recovery of the atmospheric light was presented in [22]. [32] developed a color attenuation prior and created a linear model of scene depth for the hazy image, and then learned the model parameters in a supervised way.

Deep Learning Methods: CNNs have witnessed prevailing success in computer vision tasks, and are recently introduced to haze removal. [17] exploited a multi-scale CNN (MSCNN), that first generated a coarse-scale transmission matrix and later refined it. [3] proposed a trainable *end-to-end model for medium transmission estimation*, called *DehazeNet*. It takes a hazy image as input, and outputs its transmission matrix. Combined with the global atmospheric light estimated by empirical rules, a haze-free image is recovered via the atmospheric scattering model.

All above methods share the same belief, that *in order to recover a clean scene from haze, it is the key to estimate an accurate medium transmission map*. The atmospheric light is calculated separately, and the clean image is recovered based on (1). Albeit being intuitive and physically grounded, such a procedure does not directly measure or minimize the reconstruction distortions. As a result, it will undoubtedly give rise to the sub-optimal image restoration quality. The errors in each separate estimation step will accumulate and potentially amplify each other. In contrast, AOD-Net is built with our different belief, that *the physical model could be formulated in a "more end-to-end" fashion, with all its parameters estimated in one unified model*. AOD-Net will output the dehazed clean image directly, without any intermediate step to estimate parameters. Different from [3] that performs end-to-end learning from the hazy image to the transmission matrix, the fully end-to-end formulation of AOD-Net bridges the ultimate target gap, between the hazy image and the clean image.

3. Modeling and Extension

In this section, the proposed AOD-Net is explained. We first introduce the transformed atmospheric scattering model, based on which the AOD-Net is designed. The structure of AOD-Net is then described in detail. Further, we discuss the extension of the proposed model to high-level tasks on hazy images by embedding it directly with other existing deep models, thanks to its end-to-end design.

3.1. Transformed Formula

By the atmospheric scattering model in (1), the clean image is obtained by

$$J(x) = \frac{1}{t(x)} I(x) - A\frac{1}{t(x)} + A. \tag{3}$$

As explained in Section 2, previous methods such as [17] and [3] estimate $t(x)$ and A separately and get the clean image by (3). They do not directly minimize the reconstruction errors on $J(x)$, but rather optimize the quality of $t(x)$. Such an indirect optimization causes a sub-optimal solution. Our core idea is to unify the two parameters $t(x)$ and A into one formula, i.e. $K(x)$ in (4), and directly minimize the reconstruction errors in the image pixel domain. To this end, the formula in (3) is re-expressed as

$$\begin{aligned} J(x) &= K(x)\, I(x) - K(x) + b, \text{where} \\ K(x) &= \frac{\frac{1}{t(x)}(I(x) - A) + (A - b)}{I(x) - 1}. \end{aligned} \tag{4}$$

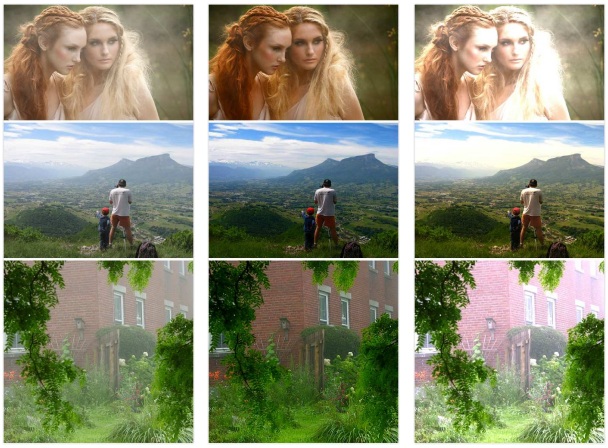
In that way, both $\frac{1}{t(x)}$ and A are integrated into the new variable $K(x)$. b is the constant bias with the default value 1. Since $K(x)$ is dependent on $I(x)$, we then aim to build an *input-adaptive* deep model, and train the model by minimizing the reconstruction errors between its output $J(x)$ and the groundtruth clean image.

A naive baseline that might be argued for is to learn $t(x)$ (or $1/t(x)$) from end to end by minimizing the reconstruction errors, with A estimated with the traditional method [8]. That requires no re-formulation of (3). To justify why jointly learning $t(x)$ and A in one is important, we compare the two solutions in experiments (see Section 4 for the synthetic settings). As observed in Figure 3, the baseline tends to overestimate A and cause overexposure visual effects. AOD-Net clearly produces more realistic lighting conditions and structural details, since the joint estimation of $\frac{1}{t(x)}$ and A enables them to mutually refine each other. In addition, the inaccurate estimate of other hyperparameters (e.g., the gamma correction), can also be compromised and compensated in the all-in-one formulation.

3.2. Network Design

The proposed AOD-Net is composed of two parts (See Figure 4): a *K-estimation module* that uses five convolutional layers to estimate $K(x)$, followed by a *clean image generation module* that consists of an element-wise multiplication layer and several element-wise addition layers to generate the recovery image via calculating (4).

The K-estimation module is the critical component of AOD-Net, being responsible for estimating the depth and relative haze level. As depicted in Figure 4 (b), we use five convolutional layers, and form multi-scale features by fusing varied size filters. [3] used parallel convolutions with varying filter sizes. [17] concatenated the coarse-scale network features with an intermediate layer of the fine-scale network. Inspired by them, the *"concat1"* layer of AOD-Net concatenates features from the layers *"conv1"* and *"conv2"*. Similarly, *"concat2"* concatenates those from *"conv2"* and *"conv3"*; *"concat3"* concatenates those from *"conv1"*, *"conv2"*, *"conv3"*, and *"conv4"*. Such a multi-scale design captures features at different scales, and the intermediate connections also compensate for the information loss during convolutions. As a simple baseline to justify concatenation, we tried on TestSetA (to be introduced in Section 4) using the structure "*"conv1"* $\rightarrow$ *"conv2"* $\rightarrow$ *"conv3"* $\rightarrow$ *"conv4"* $\rightarrow$ *"conv5"*, with no concatenation. The resulting average PSNR is 19.0674 dB and SSIM is 0.7707, both lower than current results in Table 1 (notice the large SSIM drop in particular). Notably, each convolutional layer of AOD-Net uses only three filters. As a result, our model is much light-weight, compared to existing deep methods such as [3] and [17].



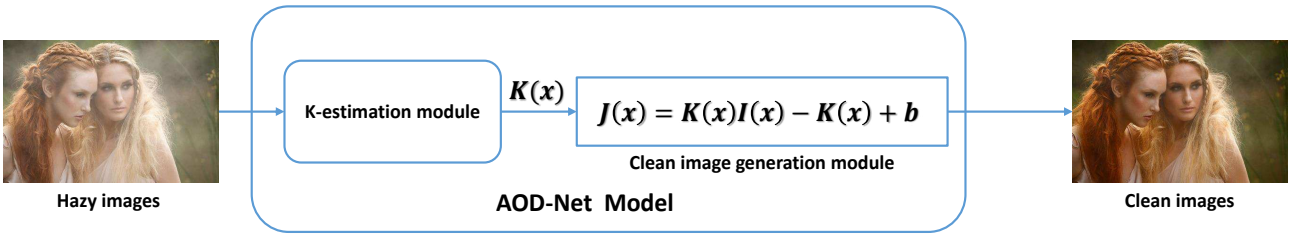
(a) Inputs (b) AOD-Net using (4) (c) Baseline using (3)

Figure 3. Visual comparison between AOD-Net using (4), and the naive baseline using (3). The images are selected from the Challenging Real Photos: see more setting details in Section 4.

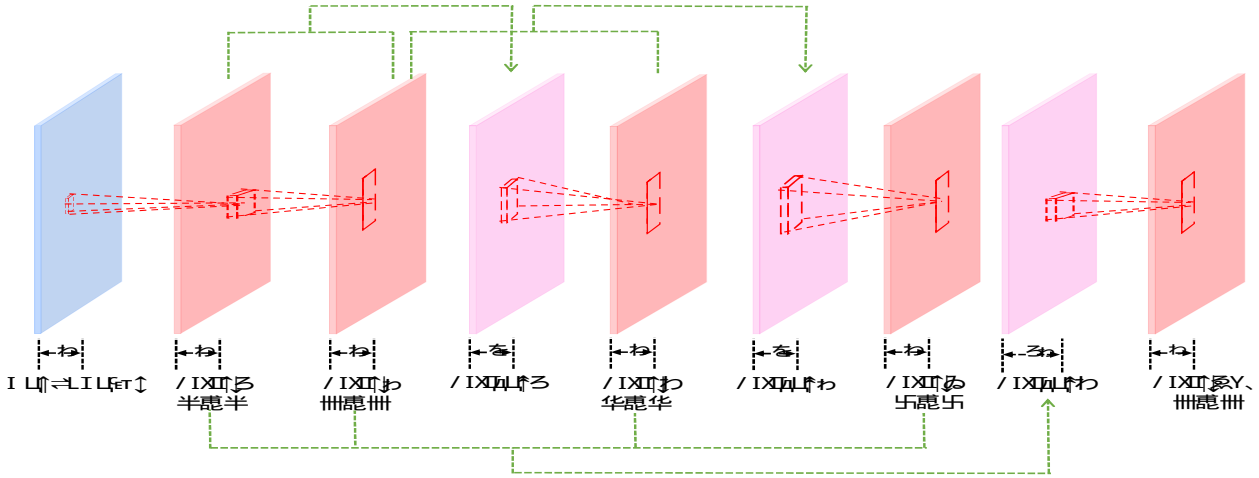
3.2.1 Necessity of K-estimation module

Most deep learning approaches for image restoration and enhancement have fully embraced *end-to-end modeling:* training a model to directly regress the clean image from the corrupted image. Examples include image denoising [29], deblurring [20], and super resolution [28]. In comparison, *there has been no end-to-end deep model for dehazing so far*[1]. While that might appear weird at the first glance, one needs to realize that haze essentially brings in non-uniform, *signal-dependent noise*: the scene attenuation of a surface caused by haze is correlated with the physical distance between the surface and the camera (i.e., the pixel depth). That is different from common image degradation models that assume *signal-independent noise*, in which case all signals go through the same parameterized degradation process. Their restoration models could thus be easily modeled

[1] [3] performed end-to-end learning from the hazy image to the transmission matrix, which is completely different from what we define here.



(a) The diagram of AOD-Net



(b) K-estimation module of AOD-Net

Figure 4. The network diagram and configuration of AOD-Net.

with one static mapping function. The same is not directly applicable to dehazing: the degradation process varies by signals, and the restoration model has to be input-adaptive as well.

3.3. Incorporation with High-Level Tasks

High-level computer vision tasks, such as object detection and recognition, concern visual semantics and have received tremendous attentions [16, 30]. However, the performance of those algorithms is largely jeopardized by various degradations. The conventional approach first resorts to a separate image restoration step as pre-processing, before feeding into the target high-level task. Recently, [27, 4] validated that a joint optimization of restoration and recognition steps would boost the performance over the traditional two-stage approach.

Previous works [31] have examined the effects and remedies for common degradations such as noise, blur and low resolution. However, to our best knowledge, there has been no similar work to quantitatively study how the existence of haze would affect high-level vision tasks, and how to alleviate its impact. Whereas current dehazing models focused merely on the restoration quality, we take the first step towards this important mission. Owing to its unique end-to-end design, AOD-Net can be seamlessly embedded with other deep models, to constitute one pipeline that performs high-level tasks on hazy images, with an implicit dehazing process. Such a pipeline can be jointly optimized from end to end for improved performance, which is infeasible if replacing AOD-Net with other deep hehazing models [3, 17].

4. Evaluations on Dehazing

4.1. Datasets and Implementation

We create *synthesized hazy images* by (1), using the ground-truth images with depth meta-data from the indoor NYU2 Depth Database [21]. We set different atmospheric lights A, by choosing each channel uniformly between $[0.6, 1.0]$, and select $\beta \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$. For the NYU2 database, we take 27, 256 images as the training set and 3,170 as the non-overlapping **TestSet A**. We also take the 800 full-size synthetic images from the Middlebury stereo database [19, 18, 9] as the **TestSet B**. Besides, we also collect a set of *natural hazy images* to evaluate our model generalization performance.

During the training process, the weights are initialized

Table 1. Average PSNR and SSIM results on TestSet A.

Metrics	ATM [22]	BCCR [12]	FVR [25]	NLD [1, 2]	DCP [8]	MSCNN [17]	DehazeNet [3]	CAP [32]	**AOD-Net**
PSNR	14.1475	15.7606	16.0362	16.7653	18.5385	19.1116	18.9613	19.6364	**19.6954**
SSIM	0.7141	0.7711	0.7452	0.7356	0.8337	0.8295	0.7753	0.8374	**0.8478**

Table 2. Average PSNR and SSIM results TestSet B.

Metrics	ATM [22]	BCCR [12]	FVR [25]	NLD [1, 2]	DCP [8]	MSCNN [17]	DehazeNet [3]	CAP [32]	**AOD-Net**
PSNR	14.3364	17.0205	16.8488	17.4479	18.9781	20.9653	21.3046	21.4544	**21.5412**
SSIM	0.7130	0.8003	0.8556	0.7463	0.8584	0.8589	0.8756	0.8879	**0.9272**

Figure 5. Visual results on dehazing synthetic images. From left to right columns: hazy images, DehazeNet results [3], MSCNN results [17], AOD-Net results, and the groundtruth images. Please amplify to view the detail differences in bounded regions.

using Gaussian random variables. We utilize ReLU neuron as we found it more effective than the BReLU neuron proposed by [3], in our specific setting. The momentum and the decay parameter are set to 0.9 and 0.0001, respectively. We adopt the simple Mean Square Error (MSE) loss function, and are pleased to find that it boosts not only PSNR, but also SSIM as well as visual quality.

The AOD-Net model takes around 10 training epochs to converge, and usually performs sufficiently well after 10 epochs. It is also found helpful to clip the gradient to constrain the norm within $[-0.1, 0.1]$. The technique has been popular in stabilizing the recurrent network training [15].

4.2. Quantitative Results on Synthetic Images

We compared the proposed model with several state-of-the-art dehazing methods: Fast Visibility Restoration (**FVR**) [25], Dark-Channel Prior (**DCP**) [8], Boundary Constrained Context Regularization (**BCCR**) [12], Automatic Atmospheric Light Recovery (**ATM**) [22], Color Attenuation Prior (**CAP**) [32], Non-local Image Dehazing (**NLD**) [1], **DehazeNet** [3], and **MSCNN** [17]. Among previous experiments, few quantitative results about the restoration quality were reported, due to the absence of haze-free ground-truth when testing on real hazy images. Our synthesized hazy images are accompanied with ground-truth images, enabling us to measure the PSNR and SSIM and to examine if the dehazed results remain faithful.

Tables 1 and 2 display the average PSNR and SSIM results on TestSets A and B, respectively. Since AOD-Net is optimized from end to end under the MSE loss, it is not surprising to see its higher PSNR performance than others. More appealing is the observation that AOD-Net obtains even greater SSIM advantages over all competitors, even though SSIM is not directly referred to as an optimization criterion. As SSIM measures beyond pixel-wise errors and is well-known to more faithfully reflect the human perception, we become curious through which part of AOD-Net, such a consistent SSIM improvement is achieved.

We conduct the following investigation: each image in TestSet B is decomposed into the sum of a mean image and a residual image. The former is constructed by all pixel locations taking the same mean value (the average 3-channel vector across the image). It is easily justified that the MSE between the two images equals the MSE between their mean images added with that between two residual images. The mean image roughly corresponds to the global illumination and is related to A, while the residual concerns more the local structural variations and contrasts, etc. We observe that AOD-Net produces the similar residual MSE (averaged on TestSet B) to a few competitive methods such as DehazeNet and CAP. However, the MSEs of the mean parts of AOD-Net results are drastically lower than DehazeNet and CAP, as shown in Table 3. Implied by that, AOD-Net could be more capable to correctly recover A (global illumination), thanks to our joint parameter estimation scheme under an end-to-end reconstruction loss. Since the human eyes are certainly more sensitive to large changes in global illumination than to any local distortion, it is no wonder why the visual results of AOD-Net are also evidently better, while some other results often look unrealistically bright.

The above advantage also manifests in the *illumination* (l) term of computing SSIM [26], and partially interprets our strong SSIM results. The other major source of SSIM gains seems to be from the *contrast* (c) term. As exam-

(a) Inputs (b) FVR (c) DCP (d) BCCR (e) ATM (f) CAP (g) NLD [1] (h) DehazeNet [3] (i) MSCNN [17] (j) AOD-Net

Figure 6. Challenging natural images results compared with the state-of-art methods.

Metrics	ATM [22]	BCCR [12]	FVR [25]	NLD [1]	DCP [8]	MSCNN [17]	DehazeNet [3]	CAP [32]	**AOD-Net**
MSE	4794.40	917.20	849.23	2130.60	664.30	329.97	424.90	356.68	**260.12**

Table 3. Average MSE between the mean images of the dehazed image and the groundtruth image, on TestSet B.

ples, we randomly select five images from TestSetB, on which the mean of *contrast* values of AOD-Net results is 0.9989, significantly higher than ATM (0.7281), BCCR (0.9574), FVR (0.9630), NLD(0.9250), DCP (0.9457) , MSCNN (0.9697), DehazeNet (0.9076), and CAP (0.9760).

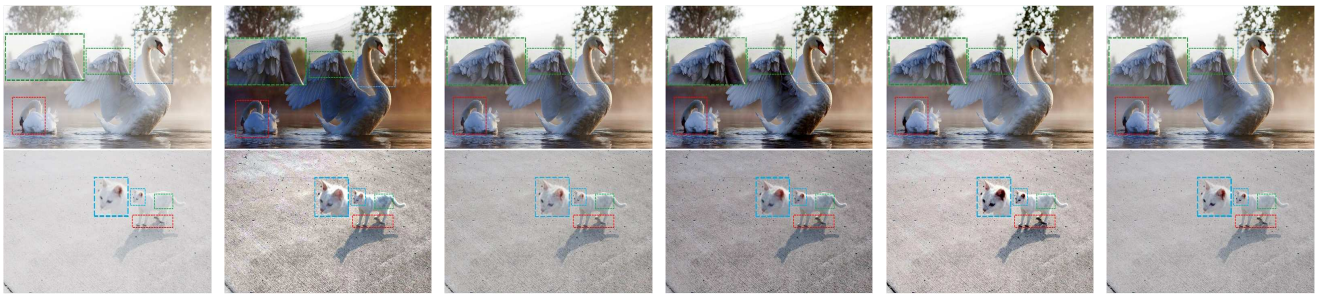
4.3. Qualitative Visual Results

Synthetic Images Figure 5 shows the dehazing results on synthetic images from TestSet A. We observe that AOD-Net results generally possess sharper contours and richer colors, and are more visually faithful to the ground-truth.

Challenging Natural Images Although trained with synthesized indoor images, AOD-Net is found to generalize well on outdoor images. We evaluate it against the state-of-the-art methods on a few natural image examples, that were found to be highly challenging to dehaze [8, 7, 3]. The challenges lie the dominance of highly cluttered objects, fine textures, or illumination variations. As revealed by Figure 6, FVR suffers from overly-enhanced visual artifacts. DCP, BCCR, ATM, NLD, and MSCNN produce unrealistic color tones on one or several images, such as DCP, BCCR and ATM results on the second row (notice the sky color), or BCCR, NLD and MSCNN results on the fourth row (notice the stone color). CAP, DehazeNet, and AOD-Net have the most competitive visual results among all, with plausible details. Yet by a closer look, we still observe that CAP sometimes blurs image textures, and DehazeNet darkens some regions. AOD-Net recovers richer and more saturated colors (compare among third- and fourth-row results), while suppressing most artifacts.

White Scenery Natural Images White scenes or object has always been a major obstacle for haze removal. Many effective priors such as [8] fail on white objects since for objects of similar color to the atmospheric light, the transmission value is close to zero. DehazeNet [3] and MSCNN [17] both rely on carefully-chosen filtering operations for post-processing, which improve their robustness to white objects but inevitably sacrifice more visual details.

Although AOD-Net does no explicitly consider the handling of white scenes, our joint optimization scheme seems



(a) Input (b) DCP [8] (c) CAP [32] (d) DehazeNet [3] (e) MSCNN [17] (f) AOD-Net

Figure 7. White scenery image dehazing results. Please amplify figures to view the detail differences in bounded regions.

Figure 8. Examples for anti-halation enhancement. Left column: real photos with halation. Right column: results by AOD-Net.

to contribute stronger robustness here. Figure 7 displays two hazy images of white scenes and their dehazing results. It is easy to notice the intolerable artifacts of DCP results, especially in the sky region of the first row. The problem is alleviated, but persists in CAP, DehazeNet and MSCNN results, while the AOD-Net results are almost artifact-free. Moreover, CAP seems to blur the textural details on white objects, while MSCNN creates the opposite artifact of over-enhancement: see the cat head region for a comparison. AOD-Net is able to remove the haze, without introducing fake color tones or distorted object contours.

Image Anti-Halation We try AOD-Net on another image enhancement task, called image anti-halation, *without re-training*. Halation is a spreading of light beyond proper boundaries, forming an undesirable fog effect in the bright areas of photos. Being related to dehazing but following different physical models, the anti-halation results by AOD-Net are decent too: see Figure 8 for a few examples.

4.4. Running Time Comparison

The light-weight structure of AOD-Net leads to faster dehazing. We select 50 images from TestSet A for all models to run, on the same machine (Intel(R) Core(TM) i7-6700 CPU@3.40GHz and 16GB memory), without GPU acceleration. The per-image average running time of all models are shown in Table 4. Despite other slower Matlab implementations, it is fair to compare DehazeNet (Pycaffe version) and ours. The results illustrate the promising efficiency of AOD-Net, costing only 1/10 time of DehazeNet per image.

Table 4. Comparison of average model running time (in seconds).

Image Size	480×640	Platform
ATM [22]	35.19	Matlab
DCP [8]	18.38	Matlab
FVR [25]	6.15	Matlab
NLD [1, 2]	6.09	Matlab
BCCR [12]	1.77	Matlab
MSCNN [17]	1.70	Matlab
CAP [32]	0.81	Matlab
DehazeNet (Matlab) [3]	1.81	Matlab
DehazeNet (Pycaffe)[2] [3]	5.09	Pycaffe
AOD-Net	**0.65**	Pycaffe

5. Improving High-level Tasks with Dehazing

We study the problem of object detection and recognition [16, 30] in the presence of haze, as an example for how high-level vision tasks can interact with dehazing. We choose the Faster R-CNN model [16] as a strong baseline[3], and test on both synthetic and natural hazy images. We then concatenate the AOD-Net model with the Faster R-CNN model, to be jointly optimized as a unified pipeline. General conclusions drawn from our experiments are: as the haze turns heavier, the object detection becomes less reliable. In all haze conditions (light, medium or heavy), our jointly tuned model constantly improves detection, surpassing both naive Faster R-CNN and non-joint approaches.

[3]We use the VGG16 model pre-trained based on 20 classes of Pascal VOC 2007 dataset provided by the Faster R-CNN authors.

Setting	Heavy + F	Heavy + A	Medium + F	Medium + A	Light + F	Light + A	Goundtruth
mAP	0.5155	0.5794	0.6046	0.6401	0.6410	0.6701	0.6990

Table 5. mAP comparison on all seven settings: "Heavy + F" and "Heavy + A" are short for "Heavy + Faster R-CNN" and "Heavy + AOD-Net followed by Faster R-CNN", respectively; similarly for the other two groups.

(a) Faster R-CNN (b) DehazeNet + Faster R-CNN (c) AOD-Net + Faster R-CNN (d) AOD-Faster R-CNN

Figure 9. Comparison of detection and recognition results on natural hazy images, using a threshold of 0.6.

Quantitative Results on Pascal-VOC 2007 with Synthetic Haze We create three synthetic sets from the Pascal VOC 2007 dataset (referred to as *Groundtruth*) [5]: *Heavy Haze* ($A = 1, \beta = 0.1$), *Medium Haze* ($A = 1, \beta = 0.06$), and *Light Haze* ($A = 1, \beta = 0.04$). The depth maps are predicted via the method described in [10]. We calculate the mean average precision (mAP) on the sets (including *Groundtruth*), using both Faster R-CNN and AOD-Net concatenated with Faster R-CNN (without joint tuning), as compared in Table 5. The heavy haze degrades mAP for nearly 0.18. By appending AOD-Net, the mAP improves by 4.54% for object detection in the light haze condition, 5.88% in the medium haze, and 12.39% in the heavy haze.

Furthermore, we jointly tune the end-to-end pipeline of AOD-Net concatenated with Faster R-CNN in the *Heavy Haze* condition, with a learning rate of 0.0001. It further boosts the mAP from 0.5794 to 0.6819, showing the impressive power of joint tuning.

Visualized Results Figure 9 displays a visual comparison of object detection results on web-source natural hazy images. Four approaches are compared: (1) *naive Faster-RCNN*: directly apply pre-trained Faster-RCNN to the hazy image; (2) *DehazeNet + Faster R-CNN*: DehazeNet concatenated with Faster R-CNN, without any joint tuning; (3) *AOD-Net + Faster R-CNN*: AOD-Net concatenated with Faster R-CNN without joint tuning; (4) *JAOD-Faster R-CNN*: jointly tuning the pipeline of AOD-Net and Faster R-CNN from end to end. We observe that haze can cause missing detections, inaccurate localizations and unconfident category recognitions for Faster R-CNN. DehazeNet tends to darken images, which often impacts detection negatively (see the first row, column (b)). While AOD-Net + Faster R-CNN already show visible advantages over naive Faster-RCNN, the performance is further dramatically improved in JAOD-Faster R-CNN results.

Note that JAOD-Faster R-CNN benefits from joint optimization in two-folds: the AOD-Net itself jointly estimates all parameters in one, and the entire pipeline tunes the low-level (dehazing) and high-level (detection and recognition) tasks from end to end.

6. Conclusion

The paper proposes AOD-Net, an all-in-one pipeline that direct reconstructs haze-free images via an end-to-end CNN. We compare AOD-Net with a variety of state-of-the-art methods, on both synthetic and natural haze images, using both objective (PSNR, SSIM) and subjective measurements. Extensive experimental results confirm the superiority, robustness, and efficiency of AOD-Net. Moreover, we present the first-of-its-kind study, on how AOD-Net can boost the object detection and recognition performance on natural hazy images, by joint tuning the pipeline.

Acknowledgement

Boyi Li and Dan Feng's research are in part supported by the National High Technology Research and Development Program (863 Program) No.2015AA015301, and the NSFC grant No. 61502191.

References

[1] D. Berman, S. Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1674–1682, 2016.

[2] D. Berman, T. Treibitz, and S. Avidan. Air-light estimation using haze-lines. In *Computational Photography (ICCP), 2017 IEEE International Conference on*, pages 1–9, 2017.

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11), 2016.

[4] S. Diamond, V. Sitzmann, S. Boyd, G. Wetzstein, and F. Heide. Dirty pixels: Optimizing image classification architectures for raw sensor data. *arXiv preprint arXiv:1701.06487*, 2017.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[6] R. Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):72, 2008.

[7] R. Fattal. Dehazing using color-lines. *ACM Transactions on Graphics (TOG)*, 34(1):13, 2014.

[8] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2011.

[9] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[10] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[11] E. J. McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York, John Wiley and Sons, Inc., 1976. 421 p.*, 1976.

[12] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision*, pages 617–624, 2013.

[13] S. G. Narasimhan and S. K. Nayar. Chromatic framework for vision in bad weather. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 598–605. IEEE, 2000.

[14] S. G. Narasimhan and S. K. Nayar. Vision and the atmosphere. *International Journal of Computer Vision*, 48(3):233–254, 2002.

[15] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *ICML*, 2013.

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[17] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *European Conference on Computer Vision*, pages 154–169. Springer, 2016.

[18] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[19] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[20] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, 2016.

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[22] M. Sulami, I. Glatzer, R. Fattal, and M. Werman. Automatic recovery of the atmospheric light in hazy images. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, pages 1–11. IEEE, 2014.

[23] R. T. Tan. Visibility in bad weather from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[24] K. Tang, J. Yang, and J. Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2995–3000, 2014.

[25] J.-P. Tarel and N. Hautiere. Fast visibility restoration from a single color or gray level image. In *Computer Vision, IEEE 12th International Conference on*, pages 2201–2208, 2009.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[27] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[28] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, and T. Huang. Self-tuned deep super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2015.

[29] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.

[30] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang. Unitbox: An advanced object detection network. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 516–520. ACM, 2016.

[31] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 770–777. IEEE, 2011.

[32] Q. Zhu, J. Mai, and L. Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, 24(11):3522–3533, 2015.