GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing

GRIDDEHAZENET: ATTENTION-BASED MULTI-SCALE NETWORK FOR IMAGE DEHAZING

BY

YONGRUI MA, B.Eng.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2019) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing
AUTHOR:	Yongrui Ma B.Eng., (Electrical and Computer Engineering) Nanjing University of Aeronautics and Astronautics, Nanjing, China
SUPERVISOR:	Dr. Jun Chen
NUMBER OF PAGES:	xiii, 51

To my dear parents and honorable supervisor

Abstract

We propose an end-to-end trainable Convolutional Neural Network (CNN), named GridDehazeNet, for single image dehazing. The GridDehazeNet consists of three modules: pre-processing, backbone, and post-processing. The trainable pre-processing module can generate learned inputs with better diversity and more pertinent features as compared to those derived inputs produced by hand-selected pre-processing methods. The backbone module implements a novel attention-based multi-scale estimation on a grid network, which can effectively alleviate the bottleneck issue often encountered in the conventional multi-scale approach. The post-processing module helps to reduce the artifacts in the final output. Experimental results indicate that the GridDehazeNet outperforms the state-of-the-art on both synthetic and real-world images. The proposed hazing method does not rely on the atmosphere scattering model, and we provide an explanation as to why it is not necessarily beneficial to take advantage of the dimension reduction offered by the atmosphere scattering model for image dehazing, even if only the dehazing results on synthetic images are concerned.

Acknowledgements

I wish to express my deepest and sincerest gratitude to all people who have helped me to successfully complete this thesis. Firstly, I am indebted to my supervisor Dr. Jun Chen, not only for his initial and ongoing support for this thesis but also for the selfless and continuous support, patience and guidance through my whole master program.

Furthermore, I would like to express my appreciation to Dr. Sorina Dumitrescu and Dr. Jiankang Zhang for being members of my defense committee. I feel grateful for their time for reviewing my thesis, and providing valuable feedbacks.

Last but not least, I am also grateful for the advice and assistance of Mr. Xiaohong Liu and Mr. Zhihao Shi, which in keeping this thesis in the schedule.

Notation and abbreviations

A	Global Atmospheric Light
ASM	Atmosphere Scattering Model
CE	Contrast Enhancement
CNN	Convolutional Neural Network
DCP	Dark Channel Prior
GC	Gamma Correction
GPU	Graphics Processing Unit
MSCNN	Multi-Scale Convolutional Neural Network
MSE	Mean Square Error
PPDN	Multi-scale Single Image Dehazing using Perceptual Pyramid Deep Network
ReLu	Rectified Linear unit
SVM	Support Vector Machine

SSIM	Structural Similarity Index
t(x)	Transmission Map
ReLU	Rectified Linear Unit
RESIDE	REalistic Single Image DEhazing dataset
WB	White Balance

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

1.1 Introduction

The image dehazing problem has received significant attention in the computer vision community over the past two decades. Image dahazing aims to recover the clear version of a hazy image. It helps mitigate the impact of image distortion induced by the environmental conditions on various visual analysis tasks, which is essential for the development of robust intelligent surveillance systems.

The Atmosphere Scattering Model (ASM) McCartney (1976); Narasimhan and Nayar (2000, 2002) provides a simple approximation of the haze effect. Specifically, it assumes that

$$I_i(x) = J_i(x)t(x) + A(1 - t(x)), \quad i = 1, 2, 3, \tag{1.1}$$

where $I_i(x)$ ($J_i(x)$) is the intensity of the ith color channel of pixel x in the hazy (clear)

(a) Hazy input (b) GridDehazeNet

Figure 1.1: Qualitative comparisons of the hazy input and our dehazed image.

image, $t(x)$ is the transmission map, and A is the global atmospheric light intensity; moreover, $t(x) = e^{-\beta d(x)}$ with β and $d(x)$ being the atmosphere scattering parameter and the scene depth, respectively. This model indicates that image dehazing is in general an underdetermined problem without the knowledge of A and $t(x)$.

As a canonical example of image restoration, the dehazing problem can be tackled using a variety of techniques that are generic in nature. Moreover, many misconceptions and difficulties encountered in image dehazing manifest in other restoration problems as well. Therefore, it is instructive to examine the relevant issues in a broader context, three of which are highlighted below.

1. Role of physical model: Many data-driven approaches to image restoration require synthetic datasets for training. To create such datasets, it is necessary to have a physical model of the relevant image degradation process (e.g., the atmosphere scattering model for the haze effect). A natural question arises whether the design of the image restoration algorithm itself should rely on this physical model. Apparently a model-dependent algorithm may suffer inherent performance loss on natural images due to model mismatch. However, it is often taken for granted that such an algorithm must have advantages on synthetic

images created using the same physical model.

2. Selection of pre-processing method: Pre-processing is widely used in image preparation to facilitate follow-up operations Tong *et al.* (2017); Ren *et al.* (2018). It can also be used to generate several variants of the given image, providing a certain form of diversity that can be harnessed via proper fusion. However, the pre-processing methods are often selected based on heuristics, thus are not necessarily best suited to the problem under consideration.

3. Bottleneck of multi-scale estimation: Image restoration requires an explicit/implicit knowledge of the statistical relationship between the distorted image and the original clear image. The statistical model needed to capture this relationship often has a huge number of parameters, comparable or even more than the available training data. As such, directly estimating these parameters based on the training data is often unreliable. Multi-scale estimation Shen *et al.* (2018); Chen *et al.* (2018) tackles this problem by i) approximating the high-dimensional statistical model by a low-dimensional one, ii) estimating the parameters of the low-dimensional model based on the training data, ii) parameterizing the neighborhood of the estimated low-dimensional model, performing a refined estimation, and repeating this procedure if needed. It is clear that the estimation accuracy on one scale will affect that on the next scale. Since multi-scale estimation is commonly done in a successive manner, its performance is often limited by a certain bottleneck.

The main contribution of this work is an end-to-end trainable CNN, named GridDehazeNet, for single image dehazing. This network can be viewed as a product of our attempt to address the aforementioned generic issues in image restoration. Firstly, the proposed GridDehazeNet does not rely on the ASM in Eq. (1.1) for haze removal, yet

capable of outperforming the existing model-dependent dehazing methods even on synthetic images; a possible explanation, together with some supporting experimental results, is provided for this puzzling phenomenon. Secondly, the pre-processing module of GridDehazeNet is fully trainable; the learned preprocessor can offer more flexible and pertinent image enhancement as compared to hand-selected pre-processing methods. Lastly, the implementation of attention-based multi-scale estimation on a grid network allows efficient information exchange across different scales and alleviate the bottleneck issue. It will be shown that the proposed dehazing method achieves superior performance in comparison with the-state-of-the-art.

1.2 Thesis Structure

To clearly introduce the advantages of the proposed GridDehazeNet, the outline of this thesis is as follow: First and foremost, Chapter 2 will make a quick review of the existing haze removal approaches; Then, Chapter 3 will introduce the proposed GridDehazeNet in detail including the model architecture, channel-wise attention and the loss function; Furthermore, Chapter 4 will make both visual and numeric comparisons between the GridDehazeNet and other state-of-the-art; Finally, Chapter 5 will make a conclusion.

Chapter 2

Background and Previous Work

2.1 Multi-image Dehazing

Early works on image dehazing either require multiple images of the same scene taken under different conditions Schechner *et al.* (2001); Shwartz *et al.* (2006); Narasimhan and Nayar (2000, 2003a); Nayar and Narasimhan (1999) or side information acquired from other sources Narasimhan and Nayar (2003b); Kopf *et al.* (2008).

2.2 Single Image Dehazing

Single image dehazing with no side information is considerably more difficult. Many methods have been proposed to address this challenge, which could be simply grouped into two categories: Prior-based Approaches and Deep Learning-based Approaches.

2.2.1 Prior-based Approaches

A conventional strategy to solve the single image haze removal problem is to estimate the transmission map $t(x)$ and the global atmospheric light A (or their variants) based on certain assumptions or priors then invert Eq. 1.1 as follow:

$$J(x) = \frac{1}{t(x)} I(x) - A\frac{1}{t(x)} + A, \tag{2.1}$$

to obtain the dehazed image. Representative works along this line of research include Tan (2008); Fattal (2008); He *et al.* (2011); Tang *et al.* (2014); Zhu *et al.* (2015).

Specifically, Tan (2008) proposes a local contrast maximization method for dehazing based on the observation that clear images tend to have higher contrast as compared to their hazy counterparts; in Fattal (2008) haze removal is realized via the analysis of albedo under the assumption that the transmission map and surface shading are locally uncorrelated.

Moreover, the dehazing method introduced in He *et al.* (2011) makes use of the Dark Channel Prior (DCP), which asserts that pixels in non-haze patches have low intensity in at least one color channel. Formally, the DCP of an image J is defined by:

$$J^{dark}(x) = \min_{c\in\{r,g,b\}} \left(\min_{y\in\Omega(x)} \left(J^c(y) \right) \right), \tag{2.2}$$

where the J^{dark} denotes the intensity of DCP, J^c is the color channel of J and $\Omega(x)$ is a local patch centered at x. And then, the transmission map $t(x)$ is estimated under mathematical deduction based on assumptions, where A is assumed to be given and $t(x)$ is

supposed to be constant for a local patch $\Omega(x)$. Taking the min of Eq. 1.1, we can obtain:

$$\min_{y\in\Omega(x)}\left(I^c(y)\right)=\hat{t}(x)\min_{y\in\Omega(x)}\left(J^c(y)\right)+(1-\hat{t}(x))A^c. \tag{2.3}$$

A^c is the color channel of A and is always supposed to be a positive number, as a result, the Eq. 2.3 is equivalent to:

$$\min_{y\in\Omega(x)}\left(\frac{I^c(y)}{A^c}\right)=\hat{t}(x)\min_{y\in\Omega(x)}\left(\frac{J^c(y)}{A^c}\right)+(1-\hat{t}(x)). \tag{2.4}$$

Then the min operation is exerted to Eq. 2.4 and we can obtain:

$$\min_c\left(\min_{y\in\Omega(x)}\left(\frac{I^c(y)}{A^c}\right)\right)=\hat{t}(x)\min_c\left(\min_{y\in\Omega(x)}\left(\frac{J^c(y)}{A^c}\right)\right)+\left(1-\hat{t}(x)\right). \tag{2.5}$$

Notice that the intensity of DCP of haze-free patches are close to 0, which means:

$$J^{dark}(x)=\min_c\left(\min_{y\in\Omega(x)}\left(J^c(x)\right)\right)=0. \tag{2.6}$$

By putting Eq. 2.6 back to Eq. 2.5, the $t(x)$ can be estimated through:

$$\hat{t}(x)=1-\min_c\left(\min_{y\in\Omega(x)}\left(\frac{I^c(y)}{A^c}\right)\right), \tag{2.7}$$

notice $\min_c\left(\min_{y\in\Omega(x)}\left(\frac{I^c(y)}{A^c}\right)\right)$ is actually the DCP intensity of normalized hazy image. To estimate A, top 0.001 brightest pixels in the dark channel are selected and then the pixels with highest intensity of I is chosen to be A. This strategy adopts information obtained from DCP, and generates a more reasonable A estimation.

Tang *et al.* (2014) suggests a machine learning approach that exploits four haze-related

features using a random forest regressor. The author finds that with the change of patch size for generating haze-related features, the extracted information would also be varied (as shown in Fig 2.1). Based on this, multi-scale DCP, multi-scale local max contrast and multi-scale local saturation are adopted to fully extract image information, where the local max contrast $C_r(x; I)$ and local saturation $S_r(x; I)$ are defined as follow:

$$C_r(x; I) = \max_{y \in \Omega_r(x)} \sqrt{\frac{1}{3\Omega_s(y)} \sum_{z \in \Omega_s(y)} \|I(z) - I(y)\|^2}, \tag{2.8}$$

$$S_r(x; I) = \max_{y \in \Omega_r(x)} \left(1 - \frac{\min_{c \in \{r,g,b\}} I^c}{\max_{c \in \{r,g,b\}} I^c}\right). \tag{2.9}$$

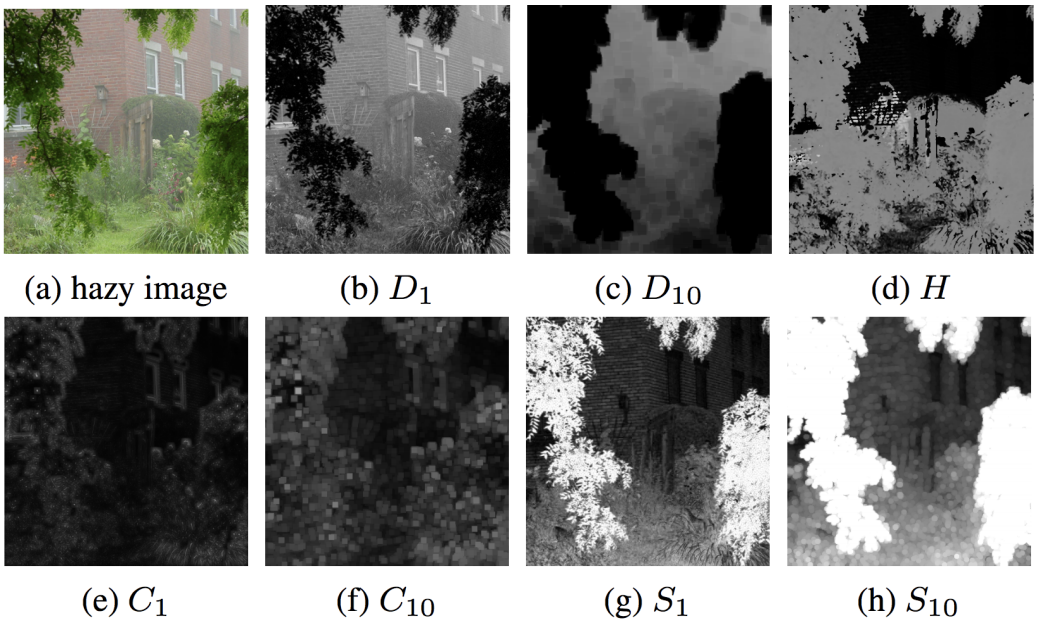
For here, $\Omega_r(y)$ denotes $r \times r$ neighbors of y, while $\Omega_s(y)$ and $|\Omega_s(y)|$ denote a $s \times s$ region and cardinality of the local neighborhood of $\Omega_s(y)$. Moreover, another prior named Hue Disparity $H(I)$ is also employed to detect haze, which could be expressed as:

$$H(I) = |I_{si}^h - I^h|, \tag{2.10}$$

where the h indicates the hue channel of the color space "Lch", and I_{si}^h is:

$$I_{si}^c = \max\left[I^c(x), 1 - I^c(x)\right], c \in \{r, g, b\}. \tag{2.11}$$

Tang *et al.* (2014) does not only provide a better transmission estimation by using random forest regressor with more abundant haze-related information, but also generate a more accurate A estimation. The mean of the intensity of I among pixels with top 0.001 largest dark channel values is leveraged to reduce the influence of noise.



(a) hazy image (b) D_1 (c) D_{10} (d) H

(e) C_1 (f) C_{10} (g) S_1 (h) S_{10}

Figure 2.1: Different feature maps for the input hazy "wall" image. (a). hazy input; (b, c). dark channel feature D_1 and D_{10}; (d). hue disparity feature; (e, f). local max contrast feature C_1 and C_{10}; (g, h) local max saturation feature S_1 and S_{10}. The subscription denotes the window size to compute the corresponding priors, like D_{10} denotes dark channel computed by 10×10 neighbors while D_1 for only the pixel itself. (Image originally used in Tang *et al.* (2014).)

The color attenuation prior is adopted in Zhu *et al.* (2015) for the development of a supervised learning method for image dehazing. Unlike all aforementioned approaches, Zhu *et al.* (2015) starts from the scene depth estimation. A, $t(x)$ and dehazed image are then generated based on estimated depth. As shown in Fig 2.2, the difference between brightness and saturation increases along the concentration of haze in an hazy image. Intuitively, this difference could be employed to estimate scene depth. Based on this, a linear model is created as follow:

$$d(x) = \theta_0 + \theta_1 v(x) + \theta_2 s(x) + \epsilon(x), \tag{2.12}$$

while v and s are brightness and saturation component of the hazy image, $\theta_0, \theta_1, \theta_2$ represents corresponding unknown coefficients and ϵ is regarded as a random variable representing the random error of the model. The coefficients of the linear model are learnt through supervised learning. What is worth mentioning is that this model has edge-preserving property as the gradient of d in Eq. 2.12 is:

$$\nabla d = \theta_1 \nabla v + \theta_2 \nabla s + \nabla \epsilon. \tag{2.13}$$

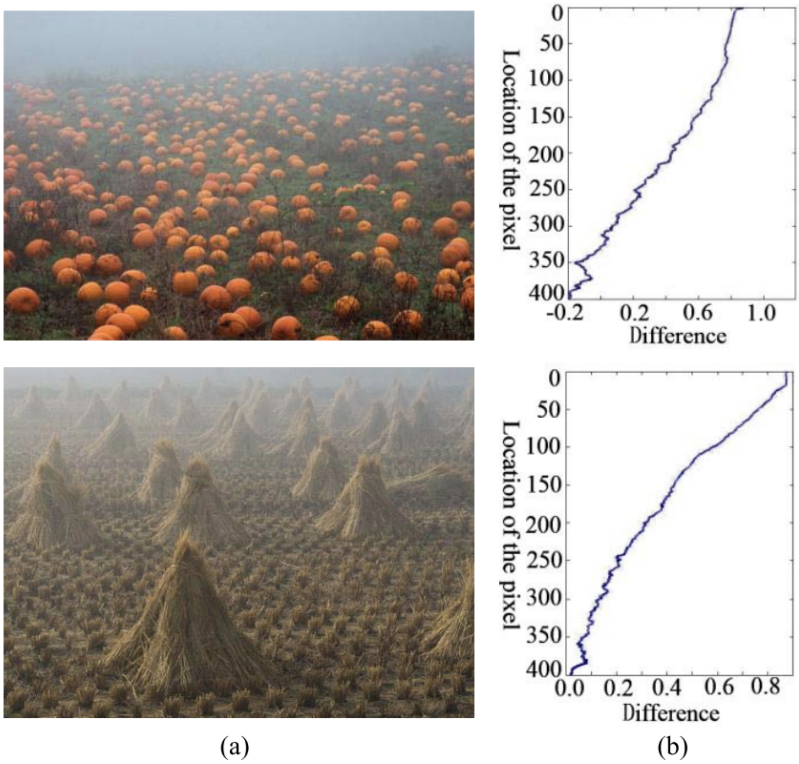


Figure 2.2: Difference between brightness and saturation increases along with the concentration of the haze. (a) A hazy image. (b) Difference between brightness and saturation. (Image originally used in Zhu *et al.* (2015).)

Although these methods have enjoyed varying degrees of success, their performances

are inherently limited by the accuracy of the adopted assumptions/priors with respect to the target scenes.

2.2.2 Deep Learning-based Approaches

With the advance in deep learning technologies and the availability of large synthetic datasets Tang *et al.* (2014), recent years have witnessed the increasing popularity of data-driven methods for image dehazing. These methods largely follow the conventional strategy mentioned above but with reduced reliance on hand-crafted priors. For example, the dehazing method, DehazeNet, proposed in Cai *et al.* (2016) uses a three-layer CNN (as shown in Fig 2.3) to directly estimate the transmission map from a given hazy image; Ren *et al.* (2016) employs a Multi-Scale CNN (MSCNN) (as shown in Fig 2.4) that is able to perform refined transmission estimation.

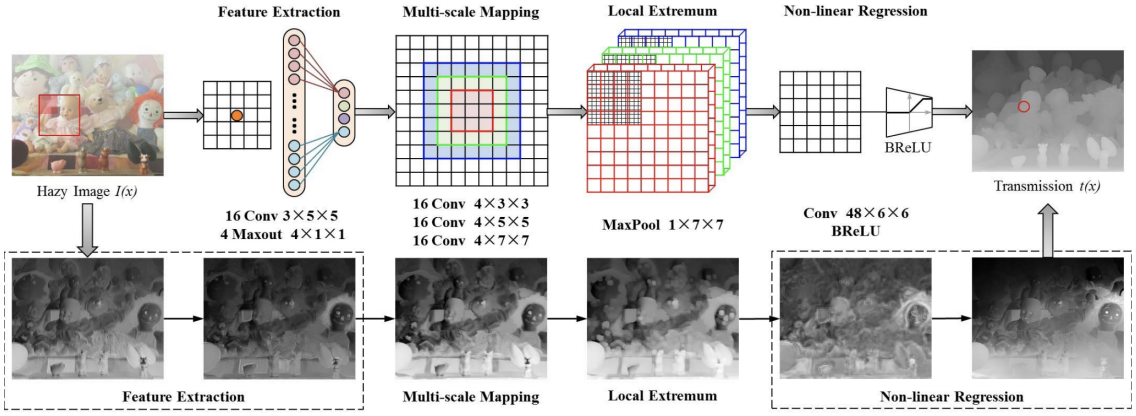


Figure 2.3: The architecture of DehazeNet. DehazeNet conceptually consists of four sequential operations (feature extraction, multi-scale mapping, local extremum and non-linear regression), which is constructed by 3 convolution layers, a max-pooling, a Maxout unit and a BReLU activation function. (Image originally used in Cai *et al.* (2016))

The AOD-Net Li *et al.* (2017) represents a departure from the conventional strategy.

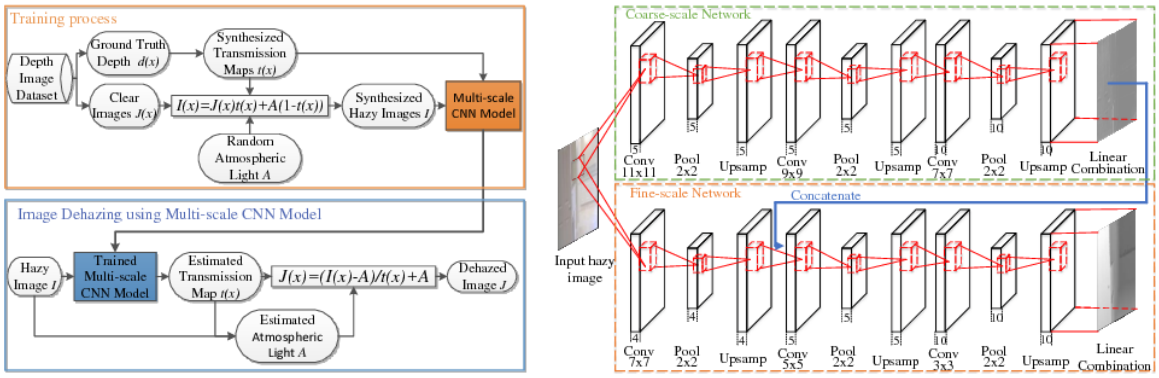


Figure 2.4: (a) Main steps of the proposed single-image dehazing algorithm. For training the multi-scale network, we synthesize hazy images and the corresponding transmission maps based on depth image dataset. In the test stage, we estimate the transmission map of the input hazy image based on the trained model, and then generate the dehazed image using the estimated atmospheric light and computed transmission map. (b) Proposed multi-scale convolutional neural network. Given a hazy image, the coarsescale network (the green dashed rectangle) predicts a holistic transmission map and feeds it to the fine-scale network (the orange dashed rectangle) in order to generate a refined transmission map. (Image originally used in Ren *et al.* (2016))

Specifically, a reformulation of Eq. 1.1 is introduced to bypass the estimation of the transmission map and the atmospheric light. To this end, the Eq. 2.1 is reformulated as:

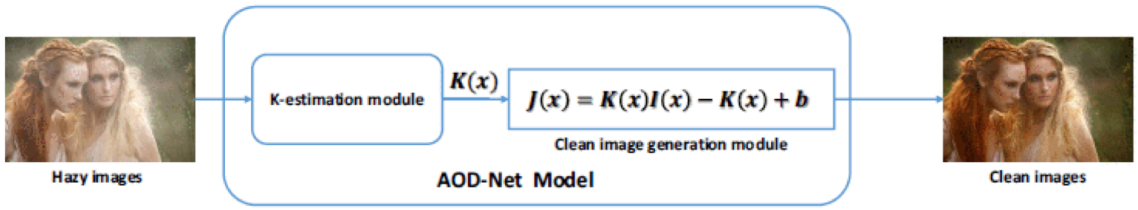
$$J(x) = K(x) I(x) - K(x) + b, \tag{2.14}$$

where

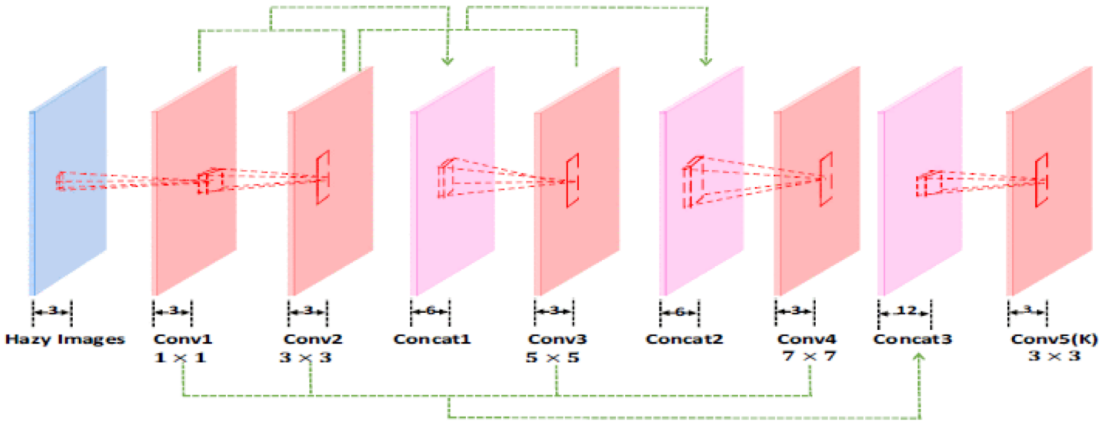
$$K(x) = \frac{\frac{1}{t(x)} (I(x) - A) + (A - b)}{I(x) - 1}. \tag{2.15}$$

In that way, both $t(x)$ and A can be integrated into $K(x)$, and b is a constant bias. As a result, the network shown in Fig 2.5 can be divided into two parts for $K(x)$ estimation and dehazed image generation based on $K(x)$. However, a close inspection reveals that

this reformulation in fact renders the ASM completely superfluous (though this point is not recognized in Li *et al.* (2017)). Ren *et al.* (2018) goes one step further by explicitly abandoning the ASM in algorithm design.



(a). The diagram of AOD-Net.



(b). K-estimation module of AOD-Net.

Figure 2.5: The diagram and configuration of AOD-Net. (Image originally used in Li *et al.* (2017))

The Gated Fusion Network (GFN) proposed in Ren *et al.* (2018) leverages hand-selected pre-processing methods (as shown in Fig 2.6) and multi-scale estimation, which are generic in nature and are subject to improvement. The selection of pre-processing approaches are observation based. The first one is that atmospheric light always changes colors of hazy image, and the second one is that the visibility of distant objects of the image might be attenuated due to scattering phenomenon. Based on these, White Balance (WB) approach is adopted to recover the latent color of the scene, while Contrast Enhancement (CE) and

Gamma Correction (GC) are employed to extract visible information. The final dehazed image is generated through weighted sum of the derived inputs, with the confidence map learnt from a multi-scale CNN shown in Fig 2.8.

Figure 2.6: Input of GFN. WB, CE and GC denote White Balance, Contrast Enhancement and Gamma Correction respectively. (Image originally used in Ren *et al.* (2018))

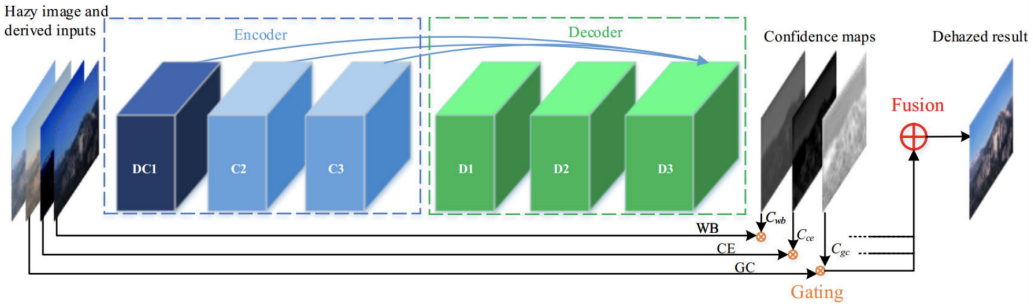


Figure 2.7: The coarsest level of GFN. The network contains layers of symmetric encoder and decoder. To expand the receptive field and extract more contextual information, dilation convolution is leveraged in the encoder block. Skip shortcuts are connected from convolutional feature maps to deconvolutional feature maps. Three enhanced versions are derived from the hazy input, and then the three inputs are weighted by three confidence maps generated by the network, respectively. (Image originally used in Ren *et al.* (2018))

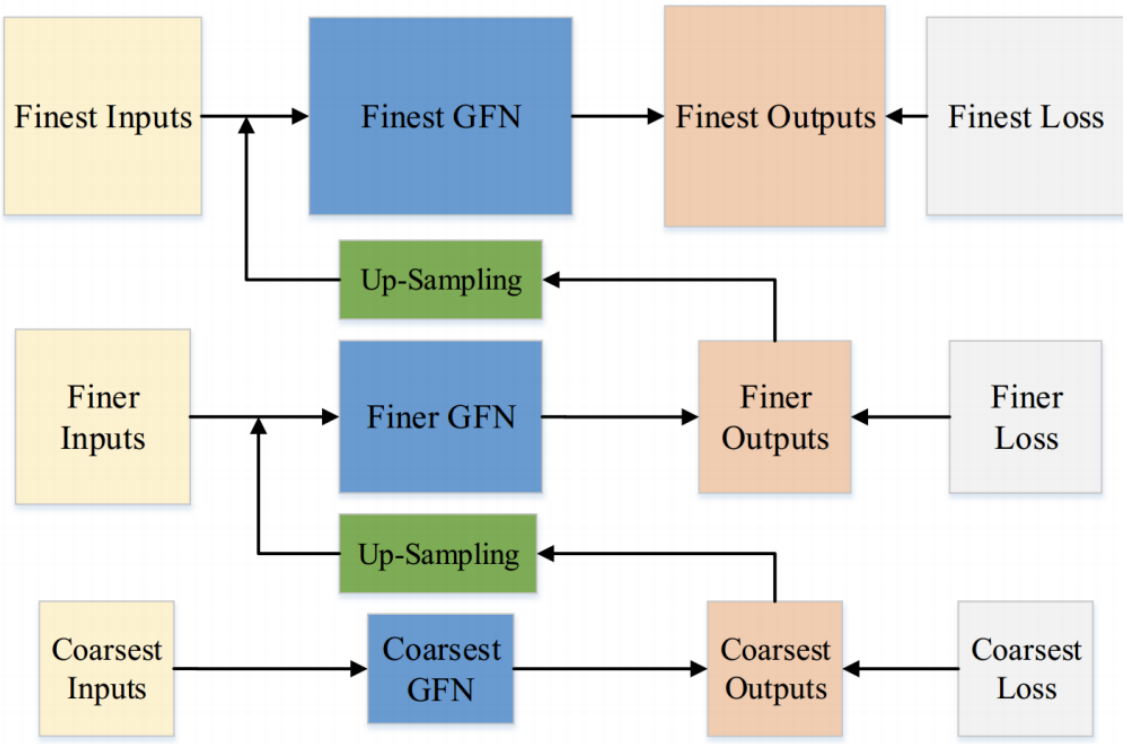


Figure 2.8: The structure of the multi-scale GFN. For each scale, the model input is the concatenation of hazy image and corresponding WB, CE and GC images. For each scale, they have a really similar structure where the coarsest level can be found in Fig 2.7. Input and corresponding ground truth are resized to fit the need of training for different scales. (Image originally used in Ren *et al.* (2018))

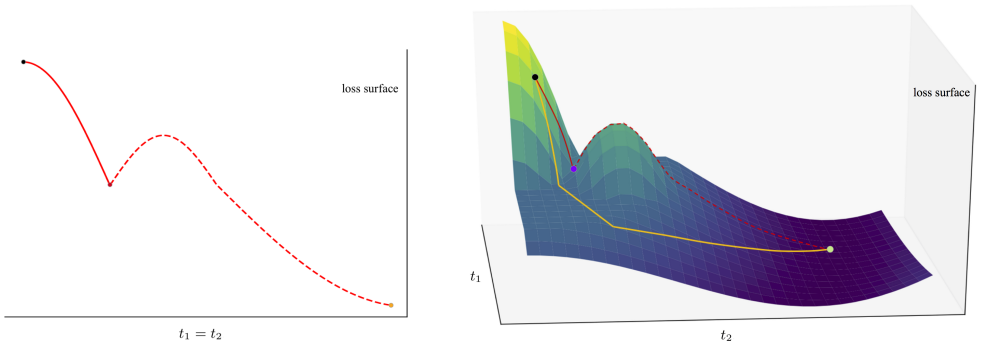
Chapter 3

GridDehazeNet

The proposed GridDehazeNet has three important features.

1. No reliance on the atmosphere scattering model: Among the aforementioned single image dehazing methods, only AOD-Net and GFN do not rely on the atmosphere scattering model. However, no convincing reason has been provided why there is any advantage in ignoring this model, as far as the dehazing results on synthetic images are concerned. The argument put forward in Ren *et al.* (2018) is that estimating $t(x)$ from a hazy image is an ill-posed problem. Nevertheless, this is puzzling since estimating $t(x)$ (which is color-channel-independent) is presumably easier than $J_i(x)$, $i = 1, 2, 3$. In Fig. **??** we offer a possible explanation why it could be problematic if one blindly uses the fact that $t(x)$ is color-channel-independent to narrow down the search space and why it might be potentially advantageous to relax this constraint in the search of the optimal $t(x)$. However, with this relaxation, the atmosphere scattering model offers no dimension reduction in the estimation procedure. More fundamentally, it is known that the loss surface of a CNN is generally well-behaved in the sense that the local minima are often almost as good as the global minimum Choromanska *et al.* (2015); Draxler *et al.* (2018); Nguyen and Hein (2018).

On the other hand, by incorporating the atmosphere scattering model into a CNN, one basically introduces a nonlinear component that is heterogeneous in nature from the rest of the network, which may create an undesirable loss surface. To support this explanation, we provide some experimental results in Section 4.6.



(a) Loss surface (b) Constrained loss surface

Figure 3.1: On the potential detrimental effect of using the atmosphere scattering model for image dehazing. For illustration purposes, we focus on two color channels of a single pixel and denote the respective transmission maps by $t1$ and $t2$. Fig. 3.1(a) plots the loss surface as a function of $t1$ and $t2$. It can be seen that the global minimum is attained a point (see the green dot) satisfying $t1 = t2$, which agrees with the ASM. With the black dot as the starting point, one can readily find this global minimum using gradient descent (see the yellow path). However, a restricted search based on the ASM along the $t1 = t2$ direction (see the red path) will get stuck at a point indicated by the purple dot (see Fig. 3.1(b)). Note that this point is a local minimum in the constrained space but not in the original space, and it becomes an obstruction simply due to the adoption of the ASM.

2. Trainable pre-processing module: The pre-processing module effectives converts the single image dehazing problem to a multi-image dehazing problem by generating several variants of the given hazy image, each highlighting a different aspect of this image and making the relevant feature information more evidently exposed. In contrast to those hand-selected pre-processing methods adopted in the existing works (e.g., Ren *et al.* (2018)), the proposed pre-processing module is made fully trainable, which is in line with the general

preference of data-driven methods over prior-based methods as shown by recent developments in image dehazing. Note that hand-selected processing methods typically aim to enhance certain concrete features that are visually recognizable. The exclusion of abstract features is not justifiable. Indeed, there might exist abstract transform domains that better suit the follow-up operations than the image domain. A trainable pre-processing module has the freedom to identify transform domains over which more diversity gain can be harnessed.

3. Attention-based multi-scale estimation: Inspired by Fourure *et al.* (2017), we implement multi-scale estimation on a grid network. The grid network has clear advantages over the conventional encoder-decoder network extensively used in image restoration Mildenhall *et al.* (2018); Zhang *et al.* (2018); Tao *et al.* (2018). In particular, the information flow in the encoder-decoder network often suffers from the bottleneck effect due to the hierarchical architecture whereas the grid network circumvents this issue via dense connections across different scales using up-sampling/down-sampling blocks. We further endow the network with a channel-wise attention mechanism, which allows for more flexible information exchange and aggregation. The attention mechanism also enables the network to better harness the diversity created by the pre-processing module.

3.1 Network Architecture

The GridDehazeNet is an end-to-end trainable network that consists of three modules, namely, the pre-processing module, the backbone module and the post-processing module. Fig. 3.2 shows the overall architecture of the proposed network.

The pre-processing module consists of a convolutional layer (w/o activation function) and a residual dense block (RDB) Zhang *et al.* (2018). It generates 16 feature maps, which

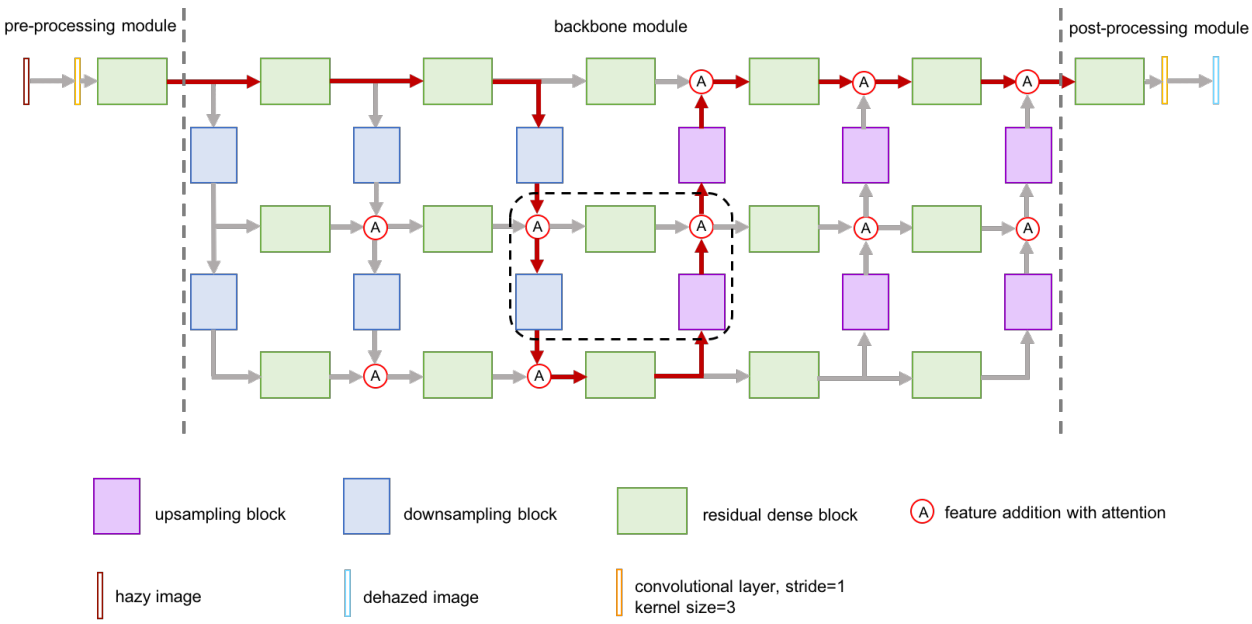


Figure 3.2: The architecture of GridDehazeNet.

will be referred to as the learned inputs, from the given hazy image.

The backbone module is an enhanced version of GridNet Fourure *et al.* (2017) originally proposed for semantic segmentation. It performs attention-based multi-scale estimation based on the learned inputs generated by the pre-processing module. In this paper, we choose a grid network with three rows and six columns. Each row corresponds to a different scale and consists of five RDB blocks that keep the number of feature maps unchanged. Each column can be regarded as a bridge that connects different scales via upsampling/downsampling modules. In each upsampling (downsampling) module, the size of feature maps is decreased (increased) by a factor of 2 while the number of feature maps is increased (decreased) by the same factor. Here upsampling/downsampling is realized using a convolutional layer instead of traditional methods such as bilinear or bicubic interpolation. Fig. 3.3 provides a detailed illustration of the RDB block, the upsampling module and the downsampling module. Each RDB block consists of five convolutional layers: the

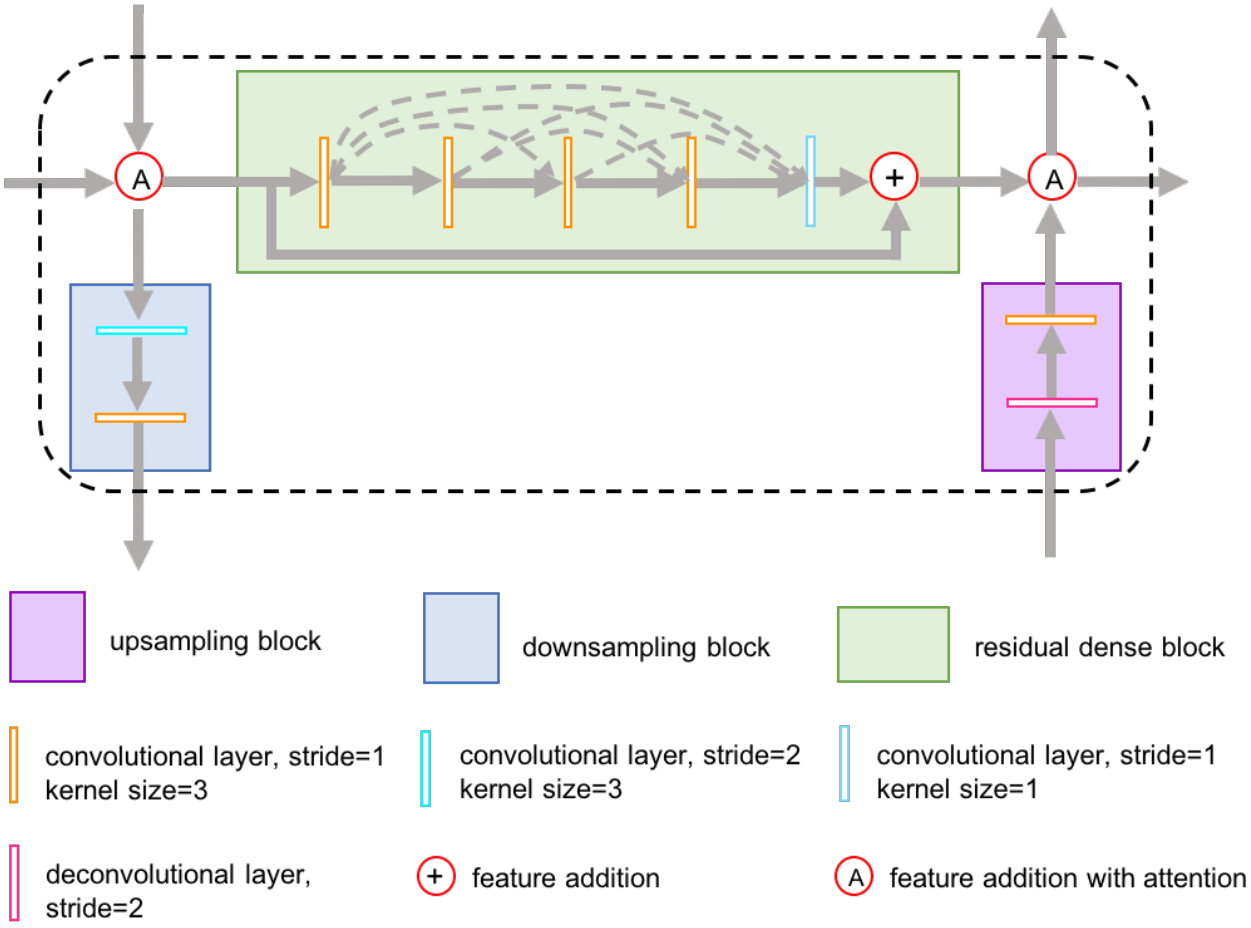


Figure 3.3: Illustration of the dash block in Fig. 3.2

first four layers are used to increase the number of feature maps while the last layer fuses these feature maps and its output is then combined with the input of this RDB block via channel-wise addition. Following Zhang *et al.* (2018), the growth rate in RDB is set to 16. The upsampling module and the downsampling module are structural the same except that different convolutional layers are used to adjust the size of feature maps. In the proposed GridDehazeNet, except for the first convolutional layer in the pre-processing module and the 1×1 convolutional layer in each RDB block, all convolutional layers employ ReLU as

the activation function. To strike a balance between the output size and the computational complexity, we set the number of feature maps at three different scales to 16, 32 and 64, respectively.

The dehazed image reconstructed directly from the output of the backbone module tends to contain artifacts. As such, we introduce a post-processing module to improve the quality of the dehazed image. The structure of the post-processing module is symmetrical to that of the pre-processing module.

It is worth mentioning that the proposed GridDehazeNet can be considered as a generalization of the classical encoder-decoder network. The red path in Fig. 3.2 illustrates an encoder-decoder structure that can be obtained by pruning the network. Moreover, with exchange branches, the proposed GridDehazeNet resembles the conventional multi-scale network.

3.2 Feature Fusion with Channel-Wise Attention

In view of the fact that feature maps from different scales may not be of the same importance, we propose a channel-wise attention mechanism, inspired by Vaswani *et al.* (2017), to generate trainable weights for feature fusion. Let F_r^i and F_c^i denote the ith feature channel from the row stream and the column stream, respectively, and let a_r^i and a_c^i denote their associated attention weights. The channel-wise attention mechanism can be expressed as

$$\tilde{F}^i = a_r^i F_r^i + a_c^i F_c^i, \tag{3.1}$$

where $\tilde{F}^i$ stands for the fused feature in the ith channel. The attention mechanism enables the GridDehazeNet to flexibly adjust the contributions from different scales in feature fusion. Our experimental results indicate that the performance of the proposed network can be greatly improved with the introduction of just a small number of trainable attention weights.

3.3 Loss Function

To train the proposed network, the smooth L_1 loss and the perceptual loss Johnson *et al.* (2016) are employed. The smooth L_1 loss provides a quantitative measure of the difference between the dehazed image and the ground truth, which is less sensitive to outliers than the MSE loss due to the fact that the L_1 norm can prevent potential gradient explosions Girshick (2015).

Smooth L_1 Loss: Let $\hat{J}_i(x)$ denote the intensity of the ith color channel of pixel x in the dehazed image, and N denote the total number of pixels. The smooth L_1 Loss can be expressed as

$$L_S = \frac{1}{N}\sum_{x=1}^{N}\sum_{i=1}^{3} F_S(\hat{J}_i(x) - J_i(x)), \tag{3.2}$$

where

$$F_S(e) = \begin{cases} 0.5e^2, & \text{if } |e| < 1, \\ |e| - 0.5, & \text{otherwise.} \end{cases} \tag{3.3}$$

Perceptual Loss: Different from the per-pixel loss, the perceptual loss leverages multi-scale features extracted from a pre-trained deep neural network to quantify the visual difference between the estimated image and the ground truth. In this paper, we use the VGG16

Simonyan and Zisserman (2014) pre-trained on ImageNet Russakovsky *et al.* (2015) as the loss network and extract the features from the last layer of each of the first three stages (i.e., Conv1-2, Conv2-2 and Conv3-3). The perceptual loss is defined as

$$L_P = \sum_{j=1}^{3} \frac{1}{C_j H_j W_j} ||\phi_j(\hat{J}) - \phi_j(J)||_2^2, \tag{3.4}$$

where $\phi_j(\hat{J})$ ($\phi_j(J)$), $j = 1, 2, 3$, denote the aforementioned three VGG16 feature maps associated with the dehazed image $\hat{J}$ (the ground truth J), and C_j, H_j and W_j specify the dimension of $\phi_j(\hat{J})$ ($\phi_j(J)$), $j = 1, 2, 3$.

Total Loss: The total loss is defined by combining the smooth L_1 loss and the perceptual loss as follows:

$$L = L_S + \lambda L_P, \tag{3.5}$$

where λ is a parameter used to adjust the relative weights on the two loss components. In this paper, λ is set to 0.04.

Chapter 4

Experimental Results

We conduct extensive experiments to demonstrate that the proposed GridDehazeNet performs favorably against the state-of-the-arts in terms of quantitative dehazing results and qualitative visual effects on synthetic and real-world datasets. The experimental results also provide useful insights into the constituent modules of GridDehazeNet and solid justifications for the overall design. The source code will be made publicly available.

4.1 Training and Testing Dataset

4.1.1 ASM-based Dataset

In general it is impossible to collect a large number of real-world hazy images and their haze-free counterparts. Therefore, data-driven dehazing methods often need to rely on synthetic hazy images, which can be generated from clear images based on the ASM via proper choice of the scattering coefficient β and the atmospheric light A. In this paper, we adopt a large-scale synthetic dataset, named RESIDE Li *et al.* (2019), to train and test the

proposed GridDehazeNet. RESIDE contains synthetic hazy images in both indoor and outdoor scenarios. The Indoor Training Set (ITS) of RESIDE contains a total of 13990 hazy indoor images, generated from 1399 clear images with $\beta \in [0.6, 1.8]$ and $A \in [0.7, 1.0]$; the depth maps $d(x)$ are obtained from the NYU Depth V2 Silberman *et al.* (2012) and Middlebury Stereo datasets Scharstein and Szeliski (2003). After data cleaning, the Outdoor Training Set (OTS) of RESIDE contains a total of 296695 hazy outdoor images, generated from 8477 clear images with $\beta \in [0.04, 0.2]$ and $A \in [0.8, 1.0]$; the depth maps of outdoor images are estimated using the algorithm developed in Liu *et al.* (2016). For testing, the Synthetic Objective Testing Set (SOTS) is adopted, which consists of 500 indoor hazy images and 500 outdoor ones. Moreover, for comparisons on real-world images, we use the dataset from Fattal (2014).

4.1.2 Camera Generated Dataset

To better prove the effectiveness of our proposed GridDehazeNet in haze removal challenge, we also test our model with an additional dataset, i.e., the NTIRE 2018 challenge on Image Dehazing Ancuti *et al.* (2018), where the hazy images are generated by adjusting the camera parameters. The NTIRE 2018 dataset includes 35 indoor images (I-HAZE) and 45 outdoor images (O-HAZE). For I-HAZE, we split the dataset into two parts, 30 for training and 5 for testing; for O-HAZE, there are 40 images in training set, and 5 in testing.

4.2 Implementation

The proposed GridDehazeNet is end-to-end trainable without the need of pre-training for sub-modules. We train the network with RGB image patches of size 240×240. For

accelerated training, the Adam optimizer Kingma and Ba (2014) is used with a batch size of 24, where β_1 and β_2 follow the default settings of 0.9 and 0.999, respectively. The initial learning rate is set to 0.001. For ITS, we train the network for 100 epochs in total and reduce the learning rate by half every 20 epochs. As for OTS, the network is trained only for 10 epochs and the learning rate is reduced by half every 2 epochs. The training is carried out on a PC with two NVIDIA GeForce GTX 1080Ti, but only one GPU is used for testing. When the training ends, the loss functions for ITS and OTS drop to 0.0005 and 0.0004, respectively, which we consider as a good indication of convergence.

4.3 Synthetic Dataset

The proposed network is tested on the synthetic dataset for qualitative and quantitative comparisons with the state-of-the-arts that include DCP He *et al.* (2011), DehazeNet Cai *et al.* (2016), MSCNN Ren *et al.* (2016), AOD-Net Li *et al.* (2017) and GFN Ren *et al.* (2018). The DCP is a prior-based method and is regarded as the baseline in single image dehazing. The others are data-driven methods. Moreover, except for AOD-Net and GFN, these methods all follow the same strategy of first estimating the transmission map and the atmosphere light then leveraging the ASM to compute the dehazed image. For fair comparisons, the above-mentioned data-driven methods are trained in the same way as the proposed one. The SOTS from RESIDE is employed as the testing dataset. We use peak signal to noise ratio (PSNR) and structure similarity (SSIM) for quantitative assessment of the dehazed outputs.

Fig. 4.1, 4.2 show the qualitative comparisons on both synthetic indoor and outdoor images from SOTS. Due to the inaccurate estimation of the haze thickness, the results of DCP are typically darker than the ground truth. Moreover, the DCP tend to cause severe

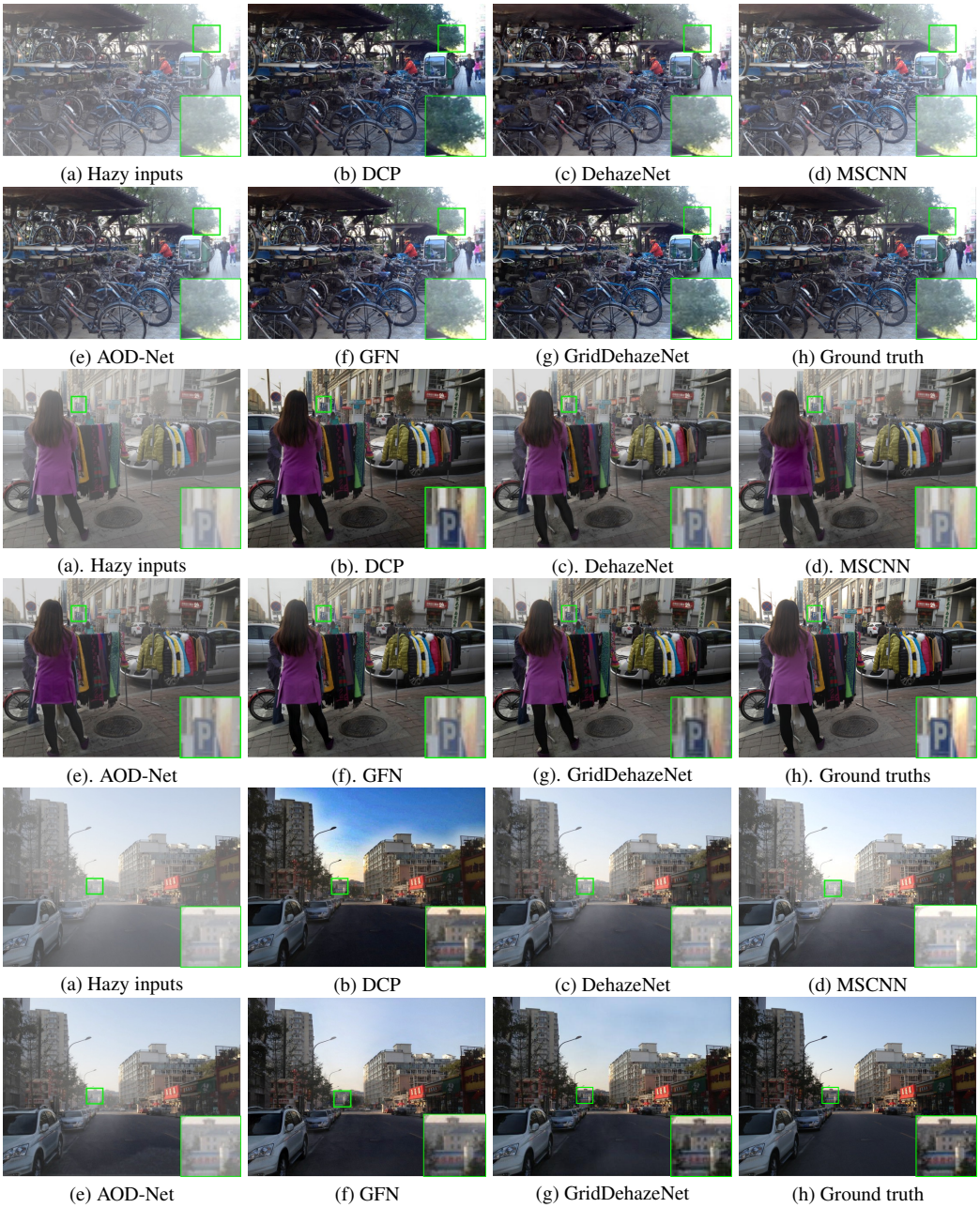
color distortions, thereby jeopardizing the quality of its output (see, *e.g.*, the tree and the sky in Fig. 4.1, 4.2 (b)). For DehazeNet as well as MSCNN, a significant amount of haze still remains unremoved and the output suffers color distortions. The AOD-Net largely overcomes the color distortion problem, but it tends to cause halo artifacts around object boundaries (see, *e.g.*, the chair leg in Fig. 4.1, 4.2 (e)) and the removal of the hazy effect is visibly incomplete. The GFN succeeds in suppressing the halo artifacts to a certain extent. However, it has limited ability to remove thick haze (see, *e.g.*, the area between two chairs and the fireplace in Fig. 4.1, 4.2 (f)). Compared with the state-of-the-arts, the proposed method has the best performance in terms of haze removal and artifact/distortion suppression (see, *e.g.*, Fig. 4.1, 4.2 (g)). The dehazed images produced by GridDehazeNet are free of major artifacts/distortions and are visually most similar to their haze-free counterparts.

Table 4.1 shows the quantitative comparisons in terms of average PSNR and SSIM values. We note that the proposed method outperforms the state-of-the-arts by a wide margin. We have also tested these dehazing methods (all pre-trained on the OTS dataset except for the DCP) directly on a new synthetic dataset. The hazy images in this new dataset are generated from 500 clear images (together with their depth maps) randomly selected from the Sun RGB-D dataset Song *et al.* (2015) through the atmosphere scattering model with $\beta \in [0.04, 0.2]$ and $A \in [0.8, 1.0]$. As shown in Table 4.1, the proposed method is fairly robust and continues to show highly competitive performance.

4.4 Real-World Dataset

We further compare the proposed method against the state-of-the-art on the real-world dataset Fattal (2014). Here we shall only make qualitative comparisons since the haze-free counterparts of the real-world hazy images in this dataset are not available. As shown

Figure 4.1: Qualitative comparisons on SOTS indoor dataset.



(a) Hazy inputs (b) DCP (c) DehazeNet (d) MSCNN

(e) AOD-Net (f) GFN (g) GridDehazeNet (h) Ground truth

(a). Hazy inputs (b). DCP (c). DehazeNet (d). MSCNN

(e). AOD-Net (f). GFN (g). GridDehazeNet (h). Ground truths

(a) Hazy inputs (b) DCP (c) DehazeNet (d) MSCNN

(e) AOD-Net (f) GFN (g) GridDehazeNet (h) Ground truth

Figure 4.2: Qualitative comparisons on SOTS outdoor dataset.

Table 4.1: Quantitative comparisons on SOTS for different methods.

Method	Indoor		Outdoor		Sun RGB-D Song *et al.* (2015)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP	16.61	0.8546	19.14	0.8605	15.81	0.8191
DehazeNet	19.74	0.8612	24.75	0.9424	23.04	0.9124
MSCNN	19.85	0.8647	22.09	0.9188	23.87	0.9262
AOD-Net	20.51	0.8516	24.14	0.9349	22.31	0.9167
GFN	24.91	0.9408	28.29	0.9731	25.35	0.9421
Proposed	**30.47**	**0.9862**	**30.05**	**0.9850**	**27.88**	**0.9658**

(a) Hazy input (b) DCP (c) DehazeNet (d) MSCNN (e) AOD-Net (f) GFN (g) Ours

Figure 4.3: Qualitative comparisons on the real-world dataset Fattal (2014).

by Fig 4.3, the results are largely consistent with those on the synthetic dataset. The DCP again suffers severe color distortions (see, *e.g.*, the sky and the girls' face in Fig 4.3 (b)). For DehazeNet, MSCNN and AOD-Net, haze removal is clearly incomplete. The GFN has limited ability to deal with dense haze and causes color distortions in some cases (see, *e.g.*, the sky and the piles in Fig 4.3 (f)). In comparison to the aforementioned methods, the proposed GridDehazeNet is more effective in haze removal and distortion suppression.

4.5 NTIRE 2018 Dataset

As shown in Table 4.2, the proposed method outperforms the state-of-the-art (note that PPDN Ancuti *et al.* (2018) is the champion of this competition and we retrain all the other methods using the same strategy adopted by PPDN).

Table 4.2: Quantitative comparisons on the NTIRE 2018 where I-HAZE (O-HAZE) stands for the indoor (outdoor) hazy dataset.

Method	I-HAZE		O-HAZE	
	PSNR	SSIM	PSNR	SSIM
DCP	12.66	0.6592	16.34	0.7480
DehazeNet	14.06	0.7293	19.31	0.8199
MSCNN	15.29	0.8087	17.40	0.8148
AOD-Net	16.38	0.8061	19.77	0.8237
GFN	21.39	0.8827	23.49	0.8782
PPDN	24.97	0.8810	24.03	0.7750
Ours	**25.85**	**0.9379**	**25.24**	**0.9352**

4.6 Atmosphere Scattering Model

To gain a better understanding of the difference between the direct estimation strategy adopted by the proposed method (where the ASM is completely bypassed) and the indirect estimation strategy (where the transmission map and the atmospheric light are first estimated, which are then leveraged to compute the dehazed image via the ASM), we repurpose the proposed GridDehazeNet for the estimation of the transmission map and the atmospheric light. Specifically, we modify the convolutional layer at the output end (i.e., the rightmost convolutional layer in Fig. 3.2) so that it outputs two feature maps, one as the estimated transmission map and the mean of the other as the estimated atmospheric light; these two estimates are then substituted into Eq. 1.1 to determine the dehazed image. The resulting network is trained in the same way as before and is tested on the SOTS. Although

adopting the ASM leads to a significant reduction in the number of parameters that need to be estimated, it in fact incurs performance degradation as shown in Table 4.3. This indicates that incorporating the ASM into the proposed network does have a detrimental effect on the loss surface.

Table 4.3: Comparisons on SOTS for different estimation strategies.

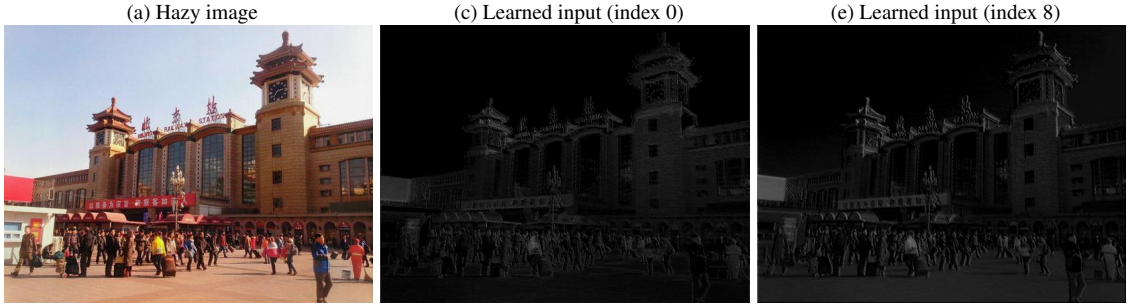
Estimation	Indoor		Outdoor	
	PSNR	SSIM	PSNR	SSIM
Indirect	28.76	0.9804	29.81	0.9837
Direct	**30.47**	**0.9862**	**30.05**	**0.9850**

4.7 Learned Inputs

Fig. 4.4 illustrates four learned inputs (out of a total of 16 learned inputs) generated by the pre-processing module. It can be seen that learned input enhances a certain aspect of the input image. For instance, the learned input with index 9 highlights a specific texture, which is not evidently shown in the given hazy image.

We conduct the following experiment to demonstrate the diversity gain offered by the learned inputs. Specifically, we remove the pre-processing module and replace the first three learned inputs by the RGB channels of the given hazy image and the rest by constant feature maps. We also conduct an experiment to show the advantages of learned inputs over those derived inputs produced by hand-selected pre-processing methods. In this case, we replace the learned inputs by the same number of derived inputs (three from the given hazy image, three from the white balanced (WB) image, three from the contrast enhanced (CE) image, three from the gamma corrected (GC) image, three from the gamma corrected GC image and one from the gray scale image). Here the use of WB, CE, GC images as derived inputs is inspired by Ren *et al.* (2018). In both cases, the resulting networks are trained

(a) Hazy image (c) Learned input (index 0) (e) Learned input (index 8)



(b) Dehazed image (d) Learned input (index 1) (f) Learned input (index 9)

Figure 4.4: Visualization of the hazy image, the dehazed image and several learned inputs.

in the same way as before and are tested on the SOTS. As shown in Table 4.4, the learned inputs offer significant diversity gain and have clear advantages over the derived inputs.

Table 4.4: Comparisons on SOTS for different types of inputs.

Input	Indoor		Outdoor	
	PSNR	SSIM	PSNR	SSIM
Original	28.93	0.9806	29.84	0.9846
Derived	29.12	0.9822	29.01	0.9790
Learned	**30.47**	**0.9862**	**30.05**	**0.9850**

4.8 Ablation Study

We perform ablation studies by considering different configurations of the backbone module of the proposed GridDehazeNet. Note that each row in the backbone module corresponds to a different scale, and the columns in the backbone module serve as bridges to

facilitate the information exchange across different scales. Table 4.5 shows how the performance and size of the proposed GridDehazeNet depends on the number of rows (denoted by r) and the number of columns (denoted by c) in the backbone module. Moreover, we also provide the qualitative comparisons for different configurations of the GridDehazeNet from Fig. 4.5-4.6. It is clear that increasing r and c leads to higher average PSNR and SSIM values. More specifically, the increase of r tends to trigger a more significant growth of model performance compared with c, while the growth of r will lead to a more dramatical increase of model size at the same time. Our full model and GFN are of of roughly the same size (225.99k vs. 212k), but the former achieves significantly better performance (see Table. 4.1). In fact, to achieve performance comparable (or even superior) to that of GFN, it suffices to use the configuration with $r = 2$ and $c = 6$ (see Table. 4.5), for which the total number of trainable weights is only 82.31k.

Table 4.5: Comparisons on SOTS for different configurations.

Configuration		Indoor		Outdoor		# Weights
		PSNR	SSIM	PSNR	SSIM	
$r = 1$	$c = 2$	22.33	0.9110	25.17	0.9475	25.35k
	$c = 4$	24.11	0.9419	26.76	0.9668	25.41k
	$c = 6$	24.63	0.9585	27.46	0.9726	25.48k
$r = 2$	$c = 2$	22.09	0.9075	25.30	0.9491	81.80k
	$c = 4$	26.19	0.9695	27.88	0.9746	82.05k
	$c = 6$	27.07	0.9756	28.08	0.9763	82.31k
$r = 3$	$c = 2$	22.10	0.9121	25.55	0.9523	224.45k
	$c = 4$	28.74	0.9812	29.38	0.9831	225.22k
	$c = 6$	**30.47**	**0.9862**	**30.05**	**0.9850**	225.99k

We perform further ablation studies by considering several variants of the proposed GridDehazeNet, which include the original GridNet Fourure *et al.* (2017), the multi-scale network resulted from removing the exchange branches (except for the first and the last

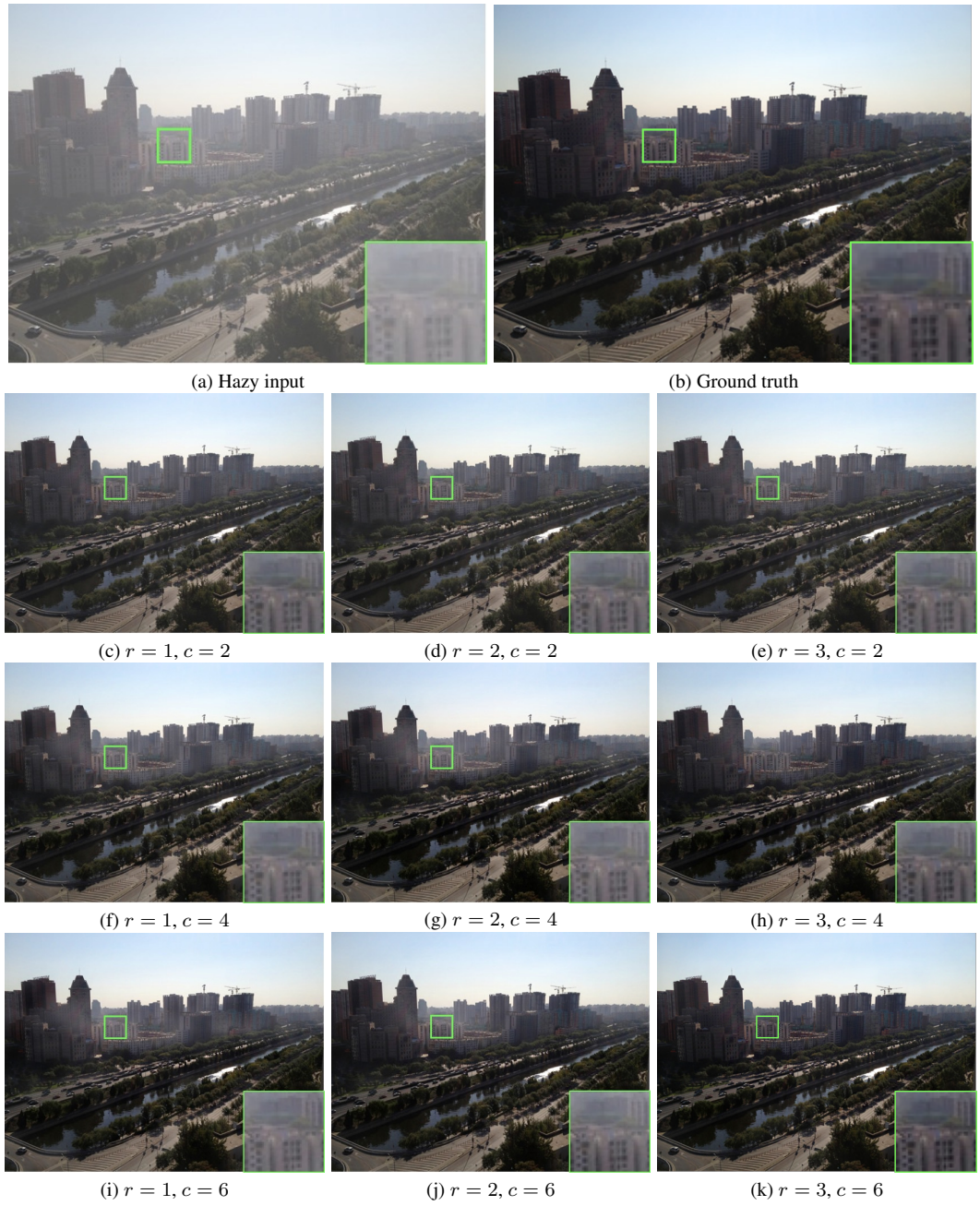
(a) Hazy input (b) Ground truth

(c) $r = 1, c = 2$ (d) $r = 2, c = 2$ (e) $r = 3, c = 2$

(f) $r = 1, c = 4$ (g) $r = 2, c = 4$ (h) $r = 3, c = 4$

(i) $r = 1, c = 6$ (j) $r = 2, c = 6$ (k) $r = 3, c = 6$

Figure 4.5: Qualitative comparisons for different configurations of GridDehazeNet.



(a) Hazy input (b) Ground truth

(c) $r = 1, c = 2$ (d) $r = 2, c = 2$ (e) $r = 3, c = 2$

(f) $r = 1, c = 4$ (g) $r = 2, c = 4$ (h) $r = 3, c = 4$

(i) $r = 1, c = 6$ (j) $r = 2, c = 6$ (k) $r = 3, c = 6$

Figure 4.6: Qualitative comparisons for different configurations of GridDehazeNet.

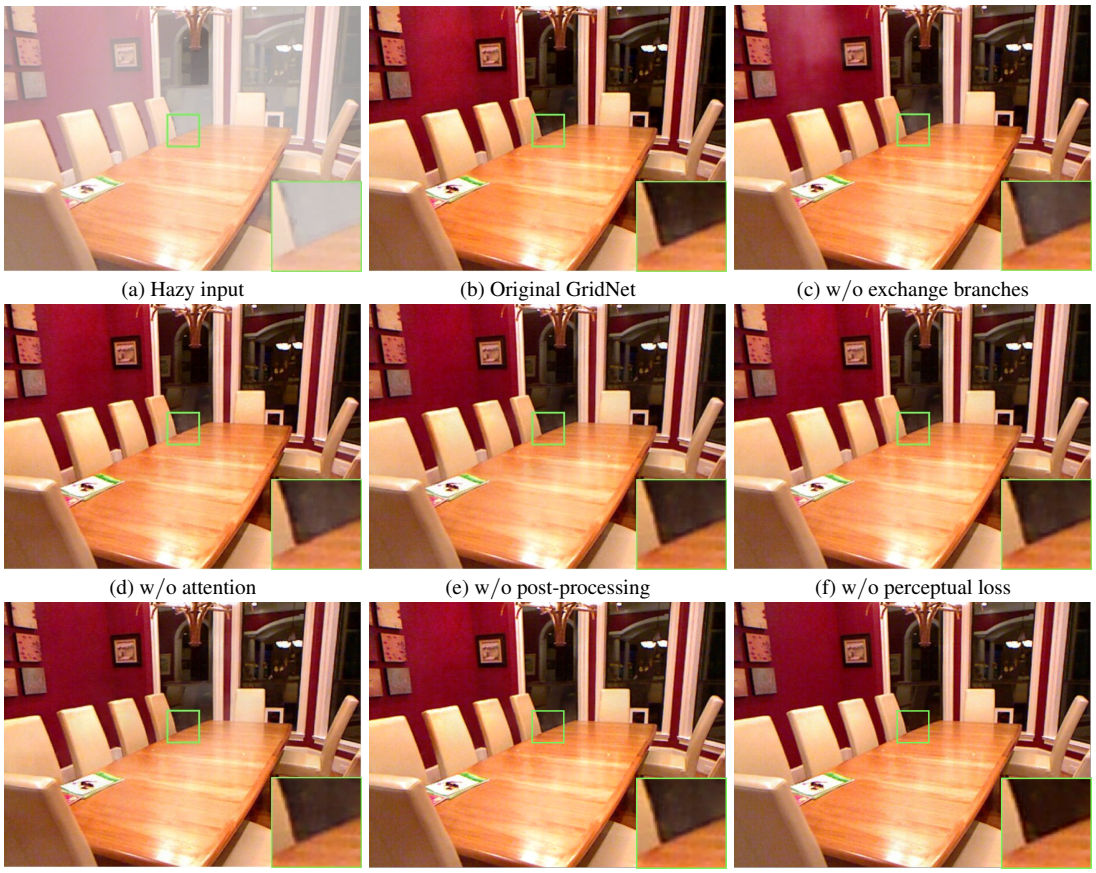
ones that are needed to maintain the minimum connection), our model without attention-based channel-wise feature fusion, without the post-processing module or without perceptual loss, as well as the encoder-decoder network obtained by pruning the proposed network. These variants are all trained in the same way as before and are tested on the SOTS. As shown in Table 4.6 and Fig. 4.7-4.8, each component has its own contribution to the performance of the full model, and the most significant improvement comes from the introduction of exchange branches, which converts the conventional multi-scale network structure to a grid network structure.

Table 4.6: Comparisons on SOTS for different variants of GridDehazeNet.

Variant	Indoor		Outdoor	
	PSNR	SSIM	PSNR	SSIM
Original GridNet Fourure *et al.* (2017)	27.37	0.9500	29.66	0.9824
w/o exchange branches	29.07	0.9672	29.65	0.9821
w/o attention	29.98	0.9784	29.82	0.9777
w/o post-processing	30.05	0.9849	29.98	0.9843
w/o perceptual loss	30.28	0.9850	29.87	0.9792
encoder-decoder	26.89	0.9725	27.99	0.9766
Our full model	**30.47**	**0.9862**	**30.05**	**0.9850**

4.9 Growth Rate

We conduct a experiment to study the influence of the growth rate in residual dense block (RDB) on the performance of GridDehazeNet. The growth rate of a RDB denotes the width of each convolutional layer of the block. The quantitative comparisons on indoor and outdoor images from SOTS in terms of average PSNR and SSIM values are shown in Table 4.7; the corresponding qualitative comparisons are demonstrated in Fig. 4.9 for indoor images and in Fig. 4.10 for outdoor images. It can be seen that as the growth rate increases, the performance of GridDehazeNet improves progressively.

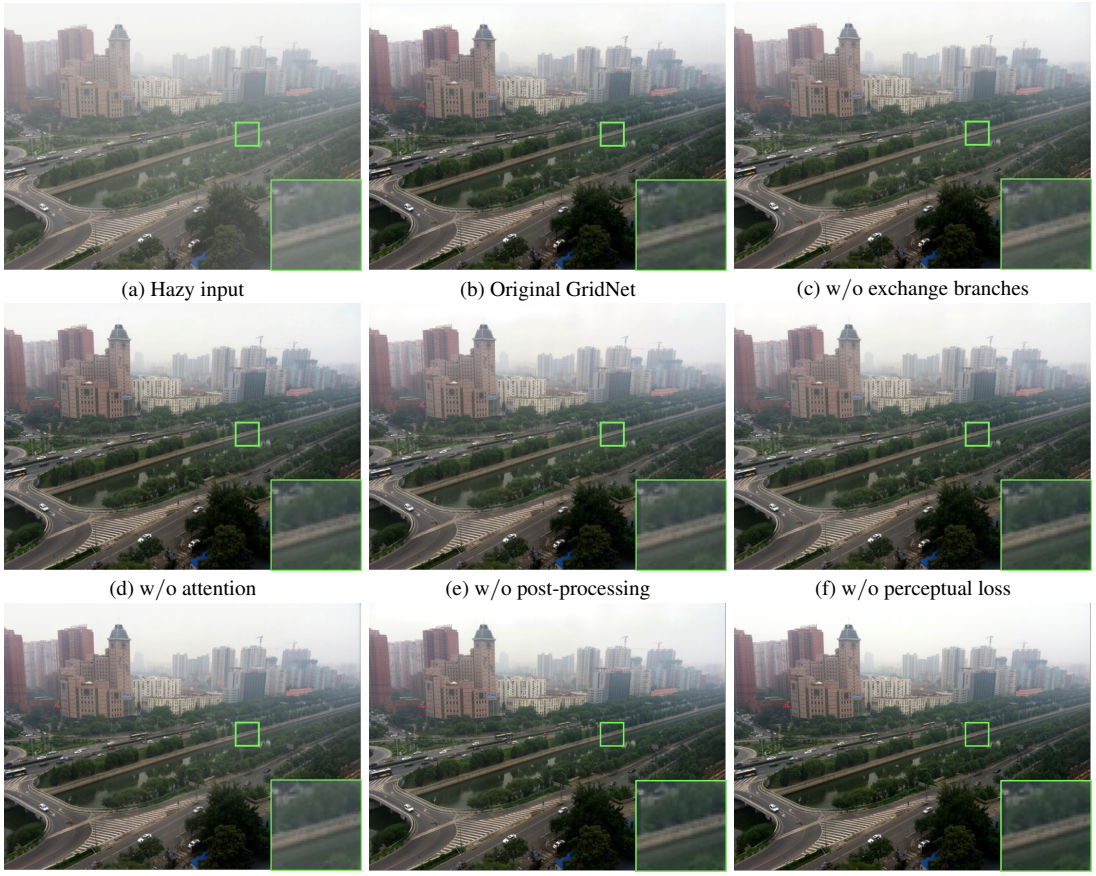


(a) Hazy input (b) Original GridNet (c) w/o exchange branches

(d) w/o attention (e) w/o post-processing (f) w/o perceptual loss

(g) encoder-decoder (h) GridDehazeNet (i) Ground truth

Figure 4.7: Qualitative comparisons for different variants of GridDehazeNet.

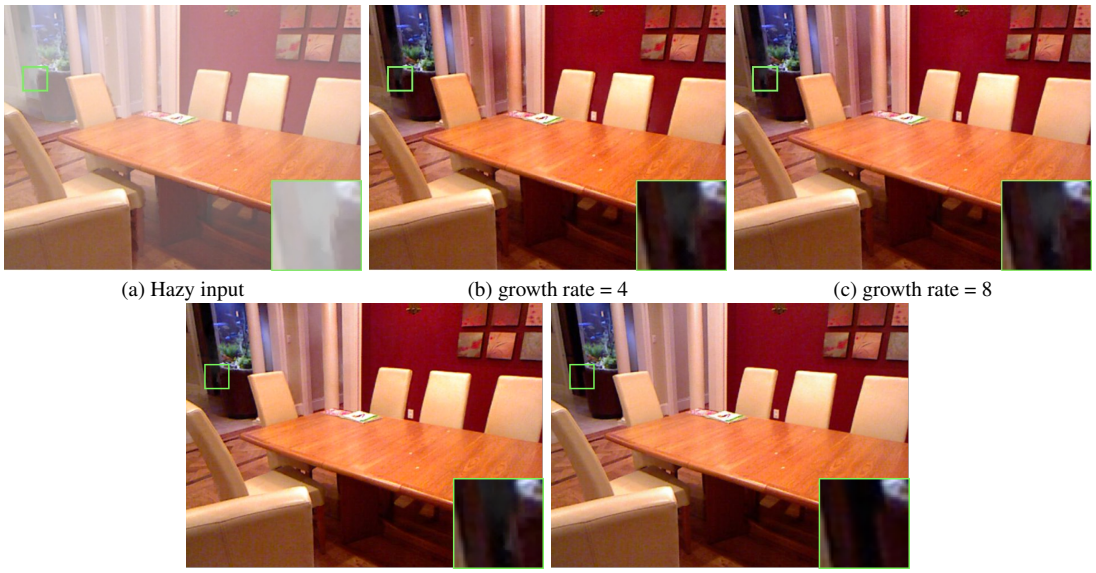


(a) Hazy input (b) Original GridNet (c) w/o exchange branches

(d) w/o attention (e) w/o post-processing (f) w/o perceptual loss

(g) encoder-decoder (h) GridDehazeNet (i) Ground truth

Figure 4.8: Qualitative comparisons for different variants of GridDehazeNet.



(a) Hazy input (b) growth rate = 4 (c) growth rate = 8

(d) GridDehazeNet (e) Ground truth

Figure 4.9: Qualitative comparisons for different growth rates.

(a) Hazy input (b) growth rate = 4 (c) growth rate = 8

(d) GridDehazeNet (e) Ground truth

Figure 4.10: Qualitative comparisons for different growth rates.

Table 4.7: Quantitative comparisons for different growth rates in terms of average PSNR and SSIM values.

Growth rate	4		8		16	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
indoor	27.77	0.9766	28.20	0.9771	**30.47**	**0.9862**
outoodr	29.16	0.9791	29.35	0.9823	**30.05**	**0.9850**

4.10 GridDehazeNet+Mask-RCNN

We shall demonstrate that the proposed GridDehazeNet can be used to improve the classification and segmentation accuracy of Mask R-CNN He *et al.* (2017) on hazy images. The experimental results are shown in Fig. 4.11-4.14. Note that more objects can be detected on the dehazed images produced by GridDehazeNet as compared to the corresponding hazy images (see, *e.g.*, the far region in Fig. 4.11-4.14). Moreover, the use of GridDehazeNet improves the confidence score of each detected object (*e.g.*, in Fig. 4.14, the confidence score of the green car in the hazy image is 0.877 whereas the corresponding score in the dehazed image is 0.933). It also leads to more accurate localization (*e.g.*, in Fig. 4.12, the green bus in the hazy image is only partially captured by the bounding box whereas it is well localized in the dehazed image) and alleviates the misclassification issue (*e.g.*, in Fig. 4.14, the bus stop is misclassfied as a train in the hazy image but not so in the dehazed image (as well as the clear image)).

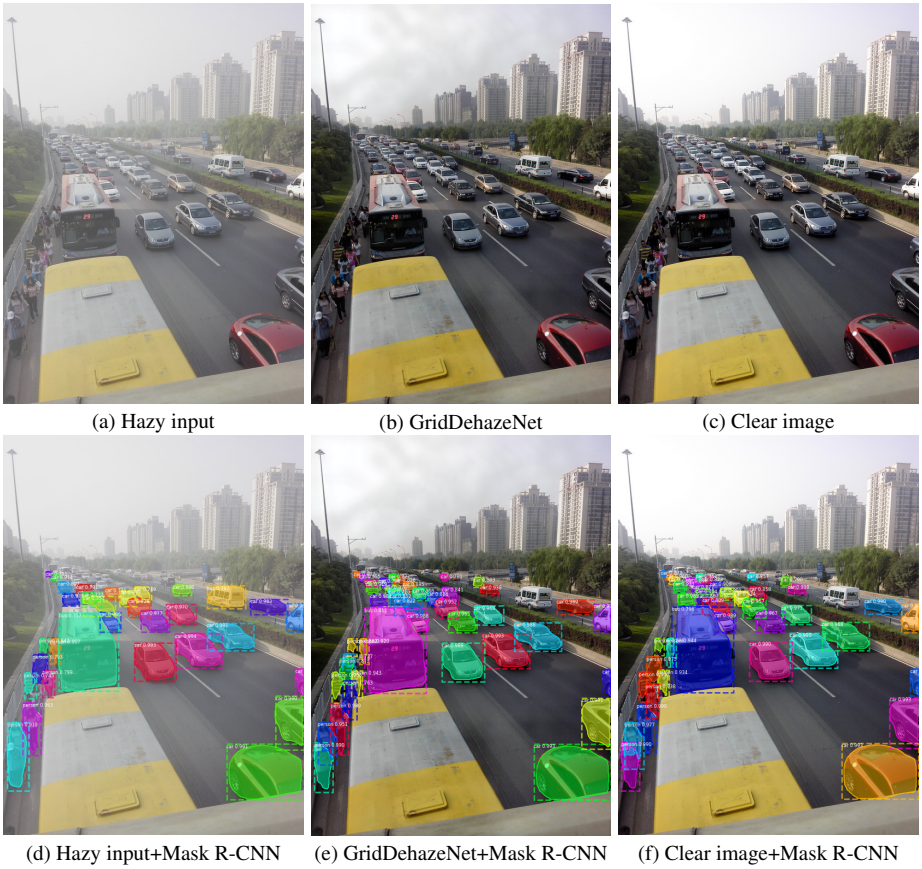
4.11 Runtime Analysis

Our un-optimized code takes about $0.22s$ to dehaze one image from SOTS on average. We have also evaluated the computational efficiency of the aforementioned state-of-the-art methods and plot their average runtimes in Fig. 4.15. It can be seen that the proposed

(a) Hazy input (b) GridDehazeNet (c) Clear image

(d) Hazy input+Mask R-CNN (e) GridDehazeNet+Mask R-CNN (f) Clear image+Mask R-CNN

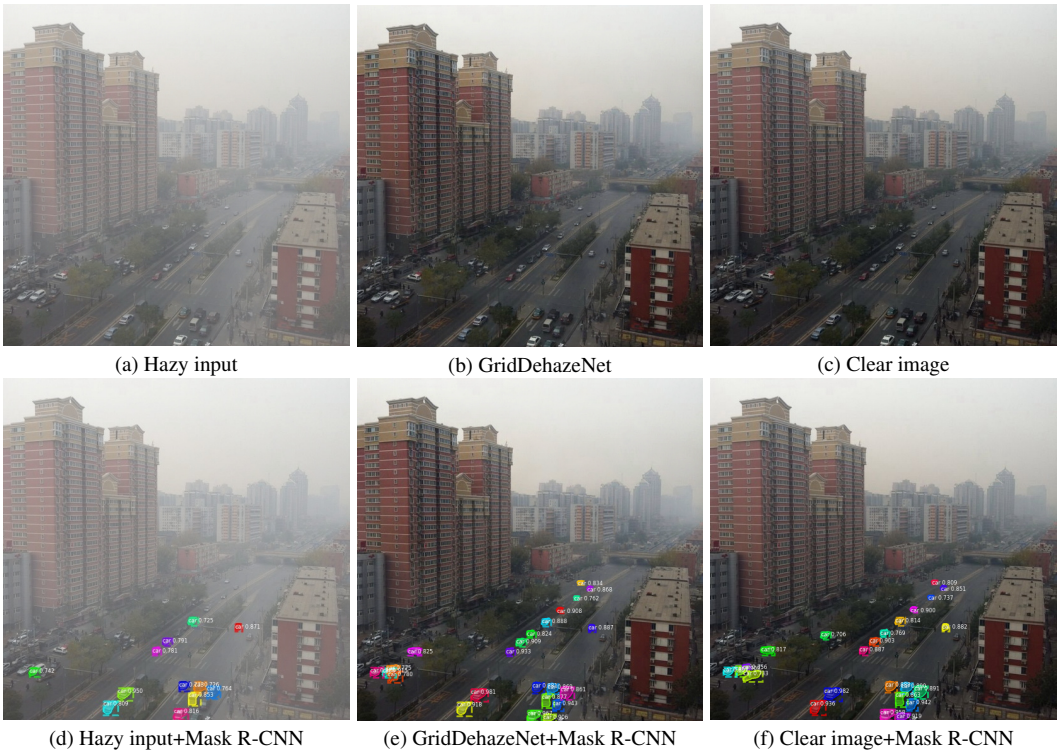
Figure 4.11: Comparisons of Mask R-CNN results on hazy, dehazed and clear images.



(a) Hazy input (b) GridDehazeNet (c) Clear image

(d) Hazy input+Mask R-CNN (e) GridDehazeNet+Mask R-CNN (f) Clear image+Mask R-CNN

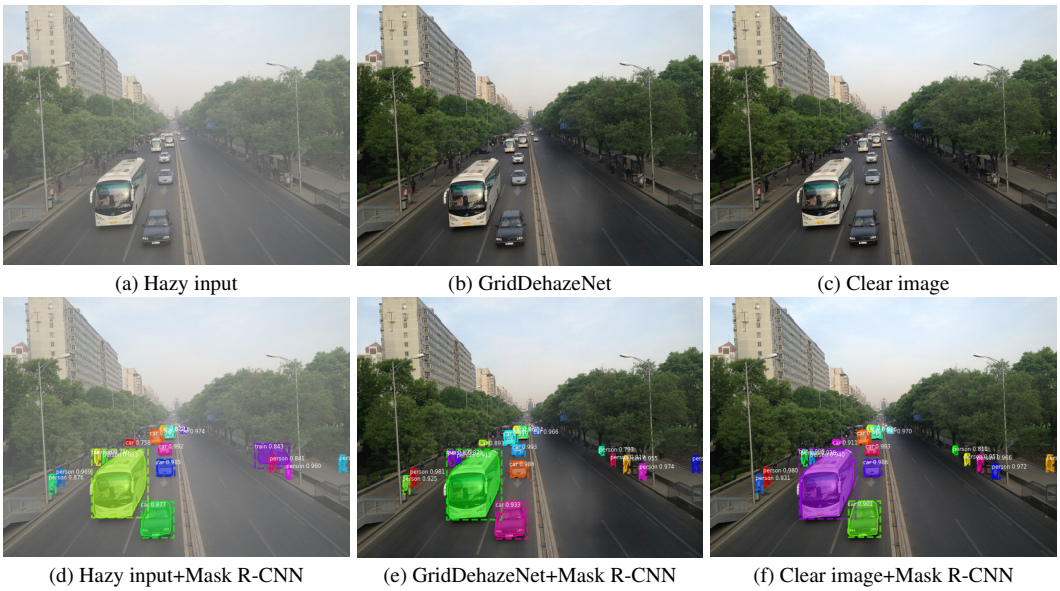
Figure 4.12: Comparisons of Mask R-CNN results on hazy, dehazed and clear images.



(a) Hazy input (b) GridDehazeNet (c) Clear image

(d) Hazy input+Mask R-CNN (e) GridDehazeNet+Mask R-CNN (f) Clear image+Mask R-CNN

Figure 4.13: Comparisons of Mask R-CNN results on hazy, dehazed and clear images.



(a) Hazy input (b) GridDehazeNet (c) Clear image

(d) Hazy input+Mask R-CNN (e) GridDehazeNet+Mask R-CNN (f) Clear image+Mask R-CNN

Figure 4.14: Comparisons of Mask R-CNN results on hazy, dehazed and clear images.

GridDehazeNet ranks second among the dehazing methods under comparison.

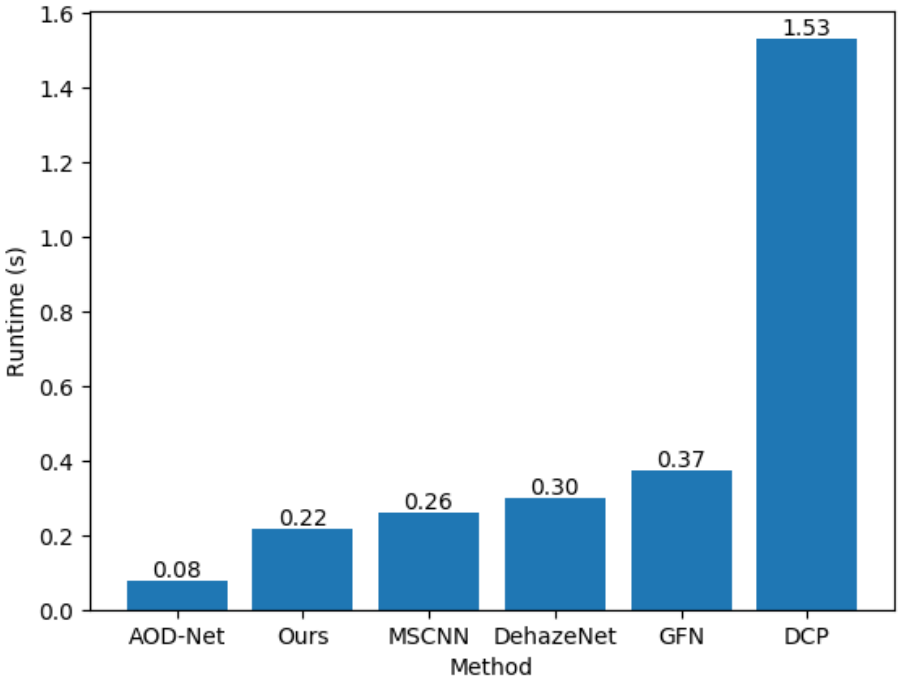


Figure 4.15: Runtime comparison of different dehazing methods.

Chapter 5

Conclusion and Future Works

We have proposed an end-to-end trainable CNN, named GridDehazeNet, and demonstrated its competitive performance for single image dehazing. Due to the generic nature of its building components, the proposed GridDehazeNet is expected to be applicable to a wide range of image restoration problems. Our work also sheds some light on the puzzling phenomenon concerning the use of the atmosphere scattering model in image dehazing, and suggests the need to rethink the role of physical model in the design of image restoration algorithms.

Bibliography

Ancuti, C., Ancuti, C. O., and Timofte, R. (2018). Ntire 2018 challenge on image dehazing: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 891–901.

Cai, B., Xu, X., Jia, K., Qing, C., and Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing (TIP)*, **25**(11), 5187–5198.

Chen, C., Chen, Q., Xu, J., and Koltun, V. (2018). Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. (2018). Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*.

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

Fattal, R. (2014). Dehazing using color-lines. *ACM transactions on graphics (TOG)*, **34**(1), 13.

Fourure, D., Emonet, R., Fromont, E., Muselet, D., Tremeau, A., and Wolf, C. (2017). Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*.

Girshick, R. (2015). Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*.

He, K., Sun, J., and Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **33**(12), 2341–2353.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., and Lischinski, D. (2008). *Deep photo: Model-based photograph enhancement and viewing*, volume 27. ACM.

Li, B., Peng, X., Wang, Z., Xu, J., and Feng, D. (2017). Aod-net: All-in-one dehazing network. In *International Conference on Computer Vision (ICCV)*, pages 4770–4778.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., and Wang, Z. (2019). Benchmarking

single-image dehazing and beyond. *IEEE Transactions on Image Processing*, **28**(1), 492–505.

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

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

Mildenhall, B., Barron, J. T., Chen, J., Sharlet, D., Ng, R., and Carroll, R. (2018). Burst denoising with kernel prediction networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2510.

Narasimhan, S. G. and Nayar, S. K. (2000). Chromatic framework for vision in bad weather. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1598. IEEE.

Narasimhan, S. G. and Nayar, S. K. (2002). Vision and the atmosphere. *International journal of computer vision*, **48**(3), 233–254.

Narasimhan, S. G. and Nayar, S. K. (2003a). Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (6), 713–724.

Narasimhan, S. G. and Nayar, S. K. (2003b). Interactive (de) weathering of an image using physical models. In *IEEE Workshop on color and photometric Methods in computer Vision*, volume 6, page 1. France.

Nayar, S. K. and Narasimhan, S. G. (1999). Vision in bad weather. In *International Conference on Computer Vision (ICCV)*, page 820. IEEE.

Nguyen, Q. and Hein, M. (2018). The loss surface and expressivity of deep convolutional neural networks.

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., and Yang, M.-H. (2016). Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision (ECCV)*, pages 154–169. Springer.

Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., and Yang, M.-H. (2018). Gated fusion network for single image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3253–3261.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.* (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, **115**(3), 211–252.

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

Schechner, Y. Y., Narasimhan, S. G., and Nayar, S. K. (2001). Instant dehazing of images using polarization. In *null*, page 325. IEEE.

Shen, Z., Lai, W.-S., Xu, T., Kautz, J., and Yang, M.-H. (2018). Deep semantic face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8260–8269.

Shwartz, S., Namer, E., and Schechner, Y. Y. (2006). Blind haze separation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1984–1991. IEEE.

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

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, S., Lichtenberg, S. P., and Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan, R. T. (2008). Visibility in bad weather from a single image.

Tang, K., Yang, J., and Wang, J. (2014). Investigating haze-relevant features in a learning framework for image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2995–3000.

Tao, X., Gao, H., Shen, X., Wang, J., and Jia, J. (2018). Scale-recurrent network for deep image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tong, T., Li, G., Liu, X., and Gao, Q. (2017). Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4807.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018). Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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