A Dual-Branch Attention Guided Context Aggregation Network for NonHomogeneous Dehazing

A Dual-Branch Attention Guided Context Aggregation Network for NonHomogeneous Dehazing

By Xiang Song, 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(2021) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE: A Dual-Branch Attention Guided Context Aggregation Network for NonHomogeneous Dehazing

AUTHOR: Xiang Song

B.Eng.(Electrical Engineering)

McMaster University

Ontario, Canada

SUPERVISOR: Dr.Jun Chen

NUMBER OF PAGES: xii, 50

To my dear family

Abstract

Image degradation arises from various environmental conditions due to the existence of aerosols such as fog, haze, and dust. These phenomena mitigate image visibility by creating color distortion, reducing contrast, and fainting object surfaces. Although the end-to-end deep learning approach has made significant progress in the field of homogeneous dehazing, the image quality of these algorithms in the context of non-homogeneous real-world images has not yet been satisfactory. We argue two main reasons that are responsible for the problem: 1) First, due to the unbalanced information processing of the high-level and low-level information in conventional dehazing algorithms, 2) due to lack of trainable data pairs. To address the above two problems, we propose a parallel dual-branch design that aims to balance the processing of high-level and low-level information, and through a method of transfer learning, utilize the small data sets to their full potential. The results from the two parallel branches are aggregated in a simple fusion tail, in which the high-level and low-level information are fused, and the final result is generated. To demonstrate the effectiveness of our proposed method, we present extensive experimental results in the thesis.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Jun Chen, for his invaluable guidance and understanding throughout my Master's program. In all things that matter, I am motivated by his enthusiasm and dedication. It is an honor for me to have him as my supervisor.

My appreciation is extended to Dr. Xiaohong Liu and Mr. Linhui Dai, for their valuable comments, advice and encouragement. Their patience, understanding, and kindness have created a friendly environment for doing research for masters students.

Especially thanks to Mr. Hafez, it was always a pleasure to spend time and talk with him over lunches. Many thanks to Moyang Wang, your cooking skills made everyday life tasty and helped me focus on my research work.

In closing, I would like to thank my family from the bottom of my heart. Without their patience and sacrifice, it would not have been possible to complete the master study during this pandemic. To my girlfriend Jiawen, thanks for making me happier than I ever imagined, and I'll try to do the same for you for the rest of my life.

Notation and abbreviations

AGCA-Net	Dual Branch Attention Guided Context Aggregation Network
KTDN	Knowledge Transfer Dehazing Network
CNN	Convolutional Neural Network
ReLu	Rectified Linear unit
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity
GPU	Graphics Processing Unit
SOTS	Synthetic Objective Testing Set
RESIDE	Realistic Single Image Dehazing Dataset
ITS	Indoor Training Set of RESIDE
NH-Haze	Non-homogeneous Haze

Contents

Abstract iv

Acknowledgements v

Notation and abbreviations vi

1 Introduction 1

1.1 Problem Background 1

1.2 Thesis Structure 6

2 Previous Work 7

2.1 Single Image Dehazing 7

2.2 Transfer Learning 15

3 Proposed Method 17

3.1 Network Architecture 17

3.1.1 Attention Guided Context Aggregation Branch 19

3.1.2 Transfer Learning Branch 21

3.1.3 Fusion Tail 23

3.1.4 Loss Function 23

4 Experiments **27**

4.1 Training and Testing Dataset 27

4.1.1 ASM-Based Dataset 28

4.1.2 Real-World Dataset 28

4.2 Evaluation Metrics 29

4.3 Training Details 30

4.4 Ablation Study 31

4.5 Comparisons with the State-of-the-art 33

4.5.1 Result on RESIDE Dataset 34

4.5.2 Result on Dense-Haze Dataset 35

4.5.3 Result on NH-Haze Dataset 36

4.5.4 Result on NH-Haze2 Dataset 37

4.5.5 Runtime Comparison 38

5 Conclusion **43**

List of Figures

1.1 Samples from our dehazing results. 2

2.1 The workflow of the supervised dehazing method. The blue arrow indicates ASM-based methods while the green arrow stands for methods that do no require ASM. 7

2.2 An overview of the proposed DCPDN image dehazing method. There are four modules in the DCPDN: 1. Dehazing via ASM. 2. Joint discriminator. 3. Pyramid densely connected transmission map estimation net. 4. Atmospheric light estimation net. We begin by estimating the transmission map based on pyramid densely-connected transmission estimation net. Next, the atmospheric light is predicted using the U-net. Finally, we estimate the dehazed image from the estimated transmission map and atmospheric light.(Image originally used in (Zhang and Patel 2018)) 10

2.3 The diagram and configuration of AOD-Net.(Image originally used in (Li et al. 2017)) 12

2.4 The architecture of GridDehazeNet.(Image originally used in (Liu et al. 2019)) 12

2.5 An overview of the proposed GCANet, which follows a basic auto-encoder structure. The encoder part of the algorithm consists of three convolution blocks, while the decoder part consists of one deconvolution block and two convolution blocks. In order to aggregate context information without grid artifacts, smoothed dilated res-blocks have been inserted between them. The gate fusion sub-network is used to fuse the features from different levels. (Image originally used in (Chen et al. 2019)) 13

2.6 This figure illustrates grid artifacts associated with dilated convolution and the proposed smoothed dilated convolution. Each of the four points in the next layer is indicated by a different color. They are associated with completely different sets of units of the previous layer, which may cause the grid artifacts to appear. On the other hand, the smoothed dilated convolution, which introduces an extra separable and shared convolution layer before the dilated convolution, adds the dependency between input units. 14

2.7 The feature fusion attention network (FFA-Net) architecture.(Image originally used in (Qin et al. 2020)) 14

2.8 Illustrative examples about transfer learning 15

3.1 A representation of our model architecture. The model consists of two branches: attention guided context aggregation(AGCA) branch and transfer learning branch. AGCA branch consists of 14 AGCA blocks. 18

3.2 Detail structure of attention guided context aggregation block. Each block consists of two AGCA layers and a skip connection to ensure a large receptive field and maximum information flow. The illustration of detailed dilated operation has been shown on the right-hand-side of AGCA layer. (Please note that the channel attention block in green color will only be used at the first layer of the AGCA block due to computation cost.) 20
3.3 Attention Module 22
3.4 Res2Net Module 24

4.1 The visual comparison of the NTIRE2021 NH-HAZE2 ablation study 33
4.2 Qualitative visual evaluation on RESIDE(ITS). 35
4.3 Qualitative visual evaluation on Dense-Haze. 39
4.4 Qualitative visual evaluation on NH-Haze. 40
4.5 Qualitative visual evaluation on NH-HAZE2. 41
4.6 Compare the runtime performance of DCP, AOD, GCA, FFA, KTDN and our methods on NH-HAZE2. 42

List of Tables

4.1 Quantitative comparison of ablation study. TLB represents the transfer learning branch, AGCA represents the 14-layers attention guided context aggregation Branch. " $\surd$ " represents the branch loaded with ImageNet pre-trained weight, " - " represents that no pre-trained weight was loaded. 32

4.2 Quantitative comparisons of RESIDE(ITS) dataset. The best results are in **bold**, and the second best are with underline. 35

4.3 Quantitative comparisons of Dense-Haze dataset. The best results are in **bold**, and the second best are with underline. 36

4.4 Quantitative comparisons of NH-HAZE 2020 dataset. The best results are in **bold**, and the second best are with underline. 37

4.5 Quantitative comparisons of NH-HAZE 2021 dataset. The best results are in **bold**, and the second best are with underline. 38

4.6 Total number of parameter of each method. 38

Chapter 1

Introduction

1.1 Problem Background

Image degradation arises from various environmental conditions due to the existence of aerosols such as fog, haze, and dust. These phenomena mitigate image visibility by creating color distortion, reducing contrast, and fainting object surfaces, which is undesirable in applications such as autonomous driving, aerial remote sensing, and video surveillance.

Image dehazing aims to recover the haze-free image from its degraded version, which has received tremendous attention in the computer vision field and the artificial intelligence community over the past several decades. It's also been considered a crucial preprocessing step to lots of high-level vision tasks such as image classification and object detection. And many dehazing methods (He et al. 2010; Li et al. 2017; Ren et al. 2016; Qin et al. 2020; Li et al. 2019; Shu et al. 2019; Engin et al. 2018) have been proposed.

Figure 1.1: Samples from our dehazing results.

The atmospheric scattering model(ASM) has been widely used to describe the formation of hazy images. This model was first introduced by McCartney (McCartney 1976). The ASM can be usually formulated as:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1.1}$$

where $I(x)$ donates the observed intensity, $J(x)$ donates the clean image, A is the global atmosphere light, and $t(x) = e^{-\beta d(x)}$ is the transmission map, with β and $d(x)$ being the atmosphere scattering parameter and the scene depth respectively.

Traditional dehazing methods (He et al. 2010; Berman, Avidan, et al. 2016;

Zhu et al. 2015) try to use hand-crafted prior such as color attenuation prior (Zhu et al. 2015), non-local prior (Berman, Avidan, et al. 2016) and DCP prior (He et al. 2010) to estimate global atmosphere light A and transmission map $t(x)$ of the ASM. These methods have shown certain success dealing with homogeneous haze scenes, but fail on most the real-world images due to complex nature of the scenes. With the hand-crafted prior, it is difficult to predict the complex haze, thus a data-driven approach that utilizes deep learning neural networks to dehaze the image appears to be the most promising method in recent years.

Since convolutional neural network(CNN) has made great progress in image restoration tasks, the dehazing algorithm (Liu et al. 2020; Dong et al. 2020; Anvari and Athitsos 2020; Fu et al. 2021; Liu et al. 2021) naturally also has a lot of related work based on CNN. This type of method can be divided into two categories. The first category is still based on the atmospheric degradation model (Zhang and Patel 2018; Li et al. 2017; Cai et al. 2016), which uses neural networks to estimate the parameters in the model. Most of the early methods are based on this idea. However, this can easily lead to accumulating errors due to model mismatch, which makes dehazing less accurate. The second type is to use the input hazy image to directly output the hazy image without using ASM model, which is often called end-to-end (Liu et al. 2019; Ren et al. 2018; Qin et al. 2020) in deep learning. Although the end-to-end approach has made significant progress in the field of homogeneous dehazing, the image quality of these algorithms in the context of non-homogeneous real-world image has not yet been satisfactory.

The two main reasons that we think are responsible for the problem are as follows:

1. Since the non-homogeneous haze in the picture is unevenly distributed, a lot of high-level details are obscured by the hazy, and a lot of low-level color information about the image is also lost. Consequently, this requires the hazing algorithm to have a larger enough receptive field to fully extract the information obscured by the haze. Thus multi-scale structures like encoder-decoder model and U-net (Ronneberger et al. 2015; Wu et al. 2020; Fu et al. 2021) model are often employed in the state-of-the-art dehazing algorithm to achieve this purpose. Though the multi-scale structure can effectively extract image features, due to the nature of down-sampling, a significant amount of low-level color information is also lost, which results in color distortion in the outputs. Therefore, we argue that the unbalanced processing of the high-level and low-level information in the non-homogeneous haze image is an important reason for the inefficient hazy removal.

2. In general, CNN-based dehazing methods require a paired training set, that is, hazy images and their clear, and an effective dehazing model should be able to learn from the mapping of a hazy image to the clear one. However, as fog/haze is a natural phenomenon, it is difficult in reality to obtain an accurate image pair that is capable of being learned by a neural network. Prior models are trained heavily on synthetic data sets, which also impacted the model performance on real-world data. As of 2020, NTIRE2020 non-homogeneous dehazing challenge (Ancuti et al. 2020b) has released 45 image pairs that can be used for training, and this year, 25 more image pairs have been released from NTIRE2021 non-homogeneous dehazing challenge (Ancuti et al. 2021). These precious photographs were taken with professional

haze generating machines and cameras systems. However, the volume of photos contained in these public data sets continues to be limited, which presents a second challenge for researchers.

To address the above two problems, we propose a dual-branch design that aims to balance the processing of high-level and low-level information, and through a method of transfer learning, utilize the small data sets to their full potential. More specifically, our network is divided into two parallel branches, each accepting the same input. One is the conventional encoder-decoder structure, which employs a multi-scale approach that allows for the effective extraction of high-level information. At the same time, as a solution to the limitation of the small dataset, we employ the transfer learning strategy in the encoder part of the branch. A pre-trained ImageNet network (Res2Net) is used in order to transfer substantial prior knowledge, allowing the decoder to learn and extract information more efficiently.

For our second branch, we give up the multiple scales structure and let the image pass through the network in full resolution. Furthermore, in order to expand the receptive field further, we also bring dilated convolution in each layer, so that this branch can obtain more low-level information. On the other hand, we have brought attention modules in the two branches to reduce performance damage caused by complex haze. Finally, the results from the last two parallel branches are aggregated in a simple fusion tail, in which the high-level and low-level information are fused together, and the final result is generated. Although our final result still falls short of the performance of state-of-the-art systems, a large number of experiments have confirmed our design to be feasible and effective. Please refer to Fig 1.1 for samples from our result.

1.2 Thesis Structure

In the next chapter, we will give a quick review of the current existing dehazing techniques and previous work. Then, Chapter 3 will introduce the proposed network in detail including model architecture, attention module and loss function. Chapter 4 will compare our proposed network with other current state-of-art dehazing algorithms. It will also discuss datasets used and various training parameters involved in training our model. Finally, Chapter 5 concludes our thesis.

Chapter 2

Previous Work

2.1 Single Image Dehazing

In the aspect of removing haze from single hazy images, dehazing algorithms can be divided into two categories: the traditional prior based methods and deep-learning-based methods (as shown in Fig 2.1).

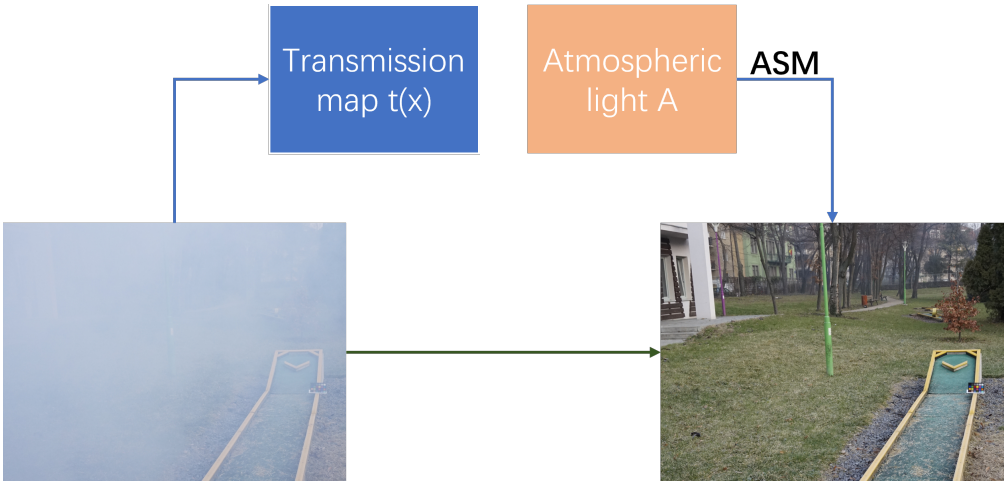

Figure 2.1: The workflow of the supervised dehazing method. The blue arrow indicates ASM-based methods while the green arrow stands for methods that do no require ASM.

Prior Based Methods: Various image priors have been proposed to remove haze from single hazy images. The prior-based methods utilize (He et al. 2010) handcrafted priors from the empirical observation to make predictions based on atmospheric scattering model. A classical way to utilize the ASM model is to estimate the unknown parameters within Eq.(1.1), where the trans-map $t(x)$ can be obtained though depth map $d(x)$, and if we can also estimate global atmospheric light A, Eq.(1.1) can be inverted as:

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

Among the process of using prior and assumption to successfully estimate J(x), (Tan 2008) observed that compared to the input hazy image, the clear image must have higher contrast. Though maximizing the contrast of local region in the dehazed image, Tan's algorithm achieved a visually pleasing result. (Zhu et al. 2014) utilized a color attenuation prior and by creating a linear model of scene depth for hazy image, supervised the learning process to learn model parameters.

Dark Channel Prior(DCP) was introduced by (He et al. 2010) which brings a more reliable way to calculate the transmission matrix. DCP asserts that hazy image may have extremely low intensities in at least one color channel. In other words, the minimum intensity of such patches should have low values. Thus we can define the DCP of the $J(x)$:

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

where J^c is the dark channel of J image, $\Omega(x)$ is a collection of areas for which the

center x is located. Base on author's assertion, the dark channel value of a natural image tends to have a very low or sometimes zero value:

$$J^{dark} = 0. \tag{2.3}$$

Considering Eq.(2.1) becomes invalid when $t(x)$ tends to be 0, a lower bound t_0 is incorporated into the Eq.(2.2):

$$J(x) = \frac{I(x) - A}{\max\left(t(x), t_0\right)} + A. \tag{2.4}$$

Suppose that the transmission map $t(x)$ is constant within local patch $\Omega(x)$, donated $\tilde{t}(x)$. A is assumed to be given, and from Eq.(1.1) we have:

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

After invoking Eq.(2.3) we have $\tilde{t}(x)$:

$$\tilde{t}(x) = 1 - \min_c \left(\min_{y \in \Omega(x)} \left(\frac{I^c(y)}{A^c} \right) \right), \tag{2.6}$$

where $\min_c \left(\min_{y \in \Omega(x)} \left(\frac{I^c(y)}{A^c} \right) \right)$ is the dark channel prior of the standardized hazy image. Now we can directly estimate the atmospheric transmission map $t(x)$.

DCP has shown to be effective for image dehazing problems. However, when the scene objects are inherently similar to sky area or no shadow appearances in the scene, DCP method tends to be invalid.

Despite the tremendous effort on utilizing image priors, all the priors still rely

on assumption and have a limit on certain target scene, leading to unpleasant results in complex real-life scenarios.

Learning Based Methods: In contrast to the above methods which rely on hand-crafted prior, deep-learning-based method (Du and Li 2018; Engin et al. 2018; Pei et al. 2019; Li et al. 2018b) is a data-driven approach that use convolutional neural network(CNN) to recover hazy images directly. For instance, (Cai et al. 2016) proposed an end-to-end model that utilizes multiple convolutional layers to estimate the medium transmission map $t(x)$ in the ASM, but they left the estimation of atmospheric light A behind.

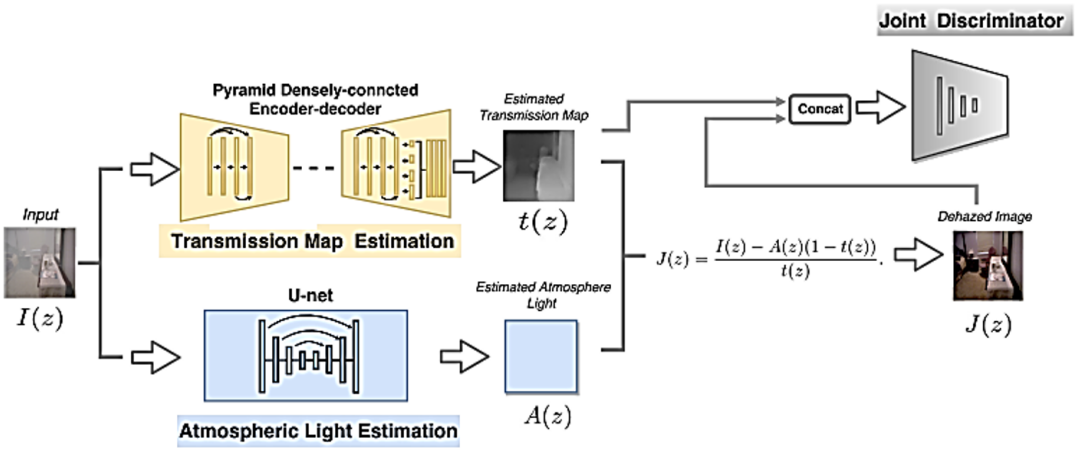

Figure 2.2: An overview of the proposed DCPDN image dehazing method. There are four modules in the DCPDN: 1. Dehazing via ASM. 2. Joint discriminator. 3. Pyramid densely connected transmission map estimation net. 4. Atmospheric light estimation net. We begin by estimating the transmission map based on pyramid densely-connected transmission estimation net. Next, the atmospheric light is predicted using the U-net. Finally, we estimate the dehazed image from the estimated transmission map and atmospheric light.(Image originally used in (Zhang and Patel 2018))

The densely connected pyramid dehazing network(as shown in Fig 2.2) is introduced by (Zhang and Patel 2018), who argue that previous method mainly focusing on the estimation of $t(x)$ without the A. So they took the advantage of U-net (Ronneberger et al. 2015) and integrated with the edge retention pyramid densely-connected encoder-decoder to better estimate $t(x)$ and A respectively.

The AOD-Net (Li et al. 2017) represents another direction by using light-weight end-to-end CNN directly to generate clear images, without estimation of transmission map and atmospheric light. This lightweight design can be embedded into other models likes Faster-RCNN. The core idea is to combine $t(x)$ and A in Eq.(1.1) as a single parameter $K(x)$; after reformulate Eq.(2.1), we get:

$$J(x) = K(x)I(x) - K(x) + c, \tag{2.7}$$

where c is a constant value. The K-estimating module in the network is responsible for estimating the $K(x)$ parameter from the input $I(x)$, followed by a clean image generation module, which uses $K(x)$ as its input adaptive parameter to estimate J (x). The network diagram can be seen from Fig 2.3.

As we mentioned earlier, that atmosphere scattering model plays an important role in the network design process of the aforementioned dehazing algorithm. This model is also widely used to produce synthetic homogeneous data set during the training process.

(Liu et al. 2019) who proposed the GridDehazeNet raises the concern that the model-dependent algorithm may perform worse on real-world images due to model

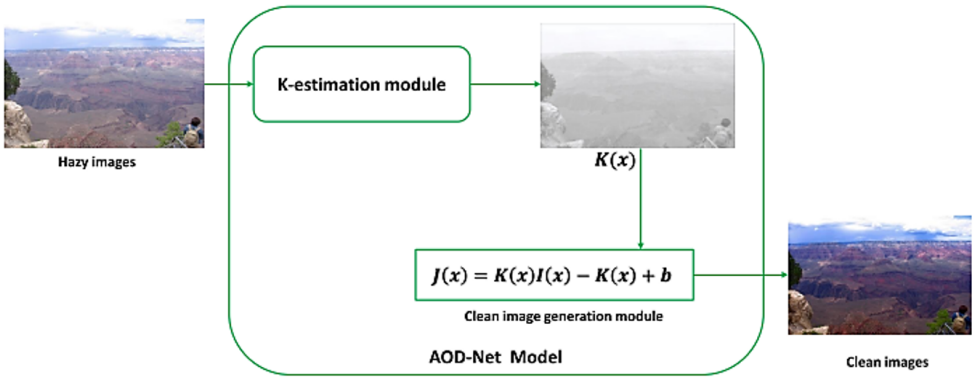

Figure 2.3: The diagram and configuration of AOD-Net.(Image originally used in (Li et al. 2017))

mismatch, thus they proposed a end-to-end method that has no reliance on the atmosphere scattering model. The success of their work sheds a light to non-model-dependent approach for dehazing tasks. Fig 2.4 illustrates their network.

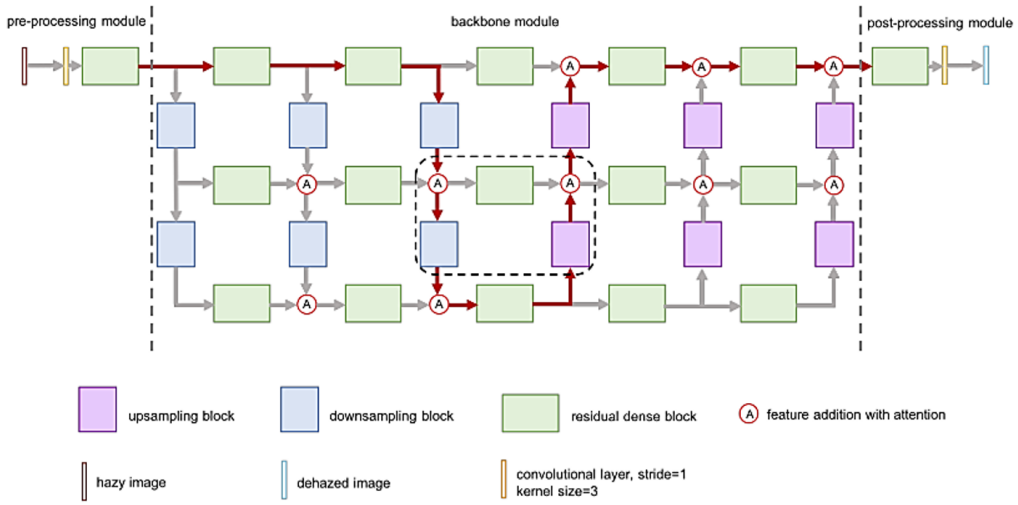

Figure 2.4: The architecture of GridDehazeNet.(Image originally used in (Liu et al. 2019))

GCA-net proposed by (Chen et al. 2019) also took a direct estimation approach, and used smoothed dilated convolution(as shown in Fig 2.5) to fix the grid artifacts

problems (as shown in Fig 2.6) introduced by conventional dilated convolution.

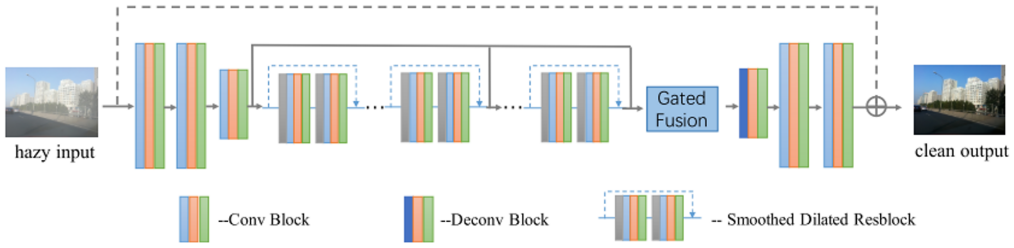

Figure 2.5: An overview of the proposed GCANet, which follows a basic auto-encoder structure. The encoder part of the algorithm consists of three convolution blocks, while the decoder part consists of one deconvolution block and two convolution blocks. In order to aggregate context information without grid artifacts, smoothed dilated res-blocks have been inserted between them. The gate fusion sub-network is used to fuse the features from different levels. (Image originally used in (Chen et al. 2019))

The gated fusion sub-network they proposed can fuse image features both on high-level and low-level coherently.

(Qin et al. 2020) proposed FFA-Network with a novel feature attention module that combines channel-wise and pixel-wise attention which achieved pleasing results. However, despite all the researching effort, there is still a domain gap between the synthetic and real world data-set for dehazing due to the difficulty of removing non-uniformly distributed haze.

To cope with the complex distribution of nonhomogeneous images, (Wu et al. 2020) introduce a knowledge transfer dehazing network(KTDN) that utilizes a teacher network as supervision to the student network, where the teacher was trained substantially using clear images only in order to provide the strong image prior for the student network to learn how a clean image should look like. The encoder part of student network is a Res2Net (Gao et al. 2019) pre-trained

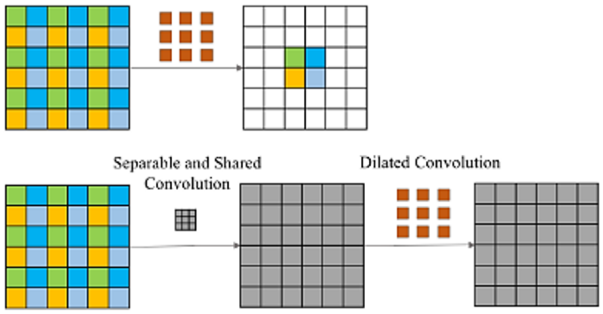

Figure 2.6: This figure illustrates grid artifacts associated with dilated convolution and the proposed smoothed dilated convolution. Each of the four points in the next layer is indicated by a different color. They are associated with completely different sets of units of the previous layer, which may cause the grid artifacts to appear. On the other hand, the smoothed dilated convolution, which introduces an extra separable and shared convolution layer before the dilated convolution, adds the dependency between input units.

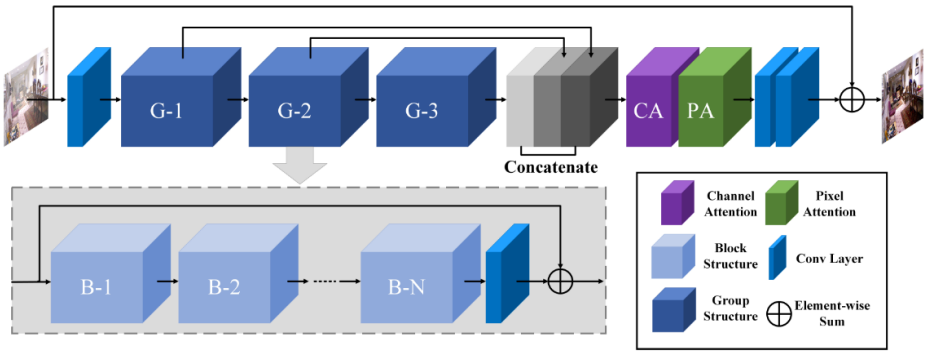

Figure 2.7: The feature fusion attention network (FFA-Net) architecture.(Image originally used in (Qin et al. 2020))

on ImageNet classification dataset. This novel approach ranks second place in NTIRE-2020 nonhomogeneous (Ancuti et al. 2020b) dehazing challenge.

2.2 Transfer Learning

Transfer learning aims at transferring knowledge between different domains to help to improve performance of the target learner. This concept may originally come from psychology research of education (Judd 1927). Psychologist Judd proposed that learning to transfer is the result of generalizing one's experiences. If a person can generalize his experiences, it is possible to transfer from one scenario to another. This process can be intuitively understood by Fig 2.8.

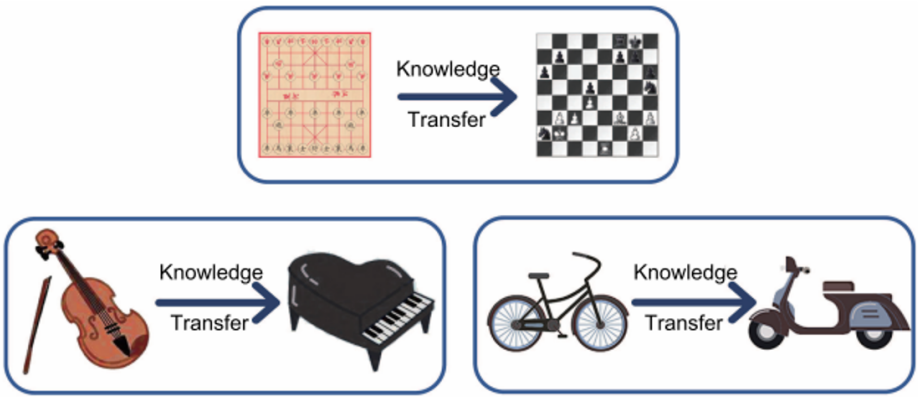

Figure 2.8: Illustrative examples about transfer learning

This concept can be also applied to machine learning tasks when there is lack of target training data. One can utilize generalized knowledge gained from another dataset to improve performance as long as two data-sets contain related source domains.

In our experiment, we utilize a network pre-trained for image classification tasks and leverage the strong feature extraction ability generalized during the process. Compared with the random initialized network, the pre-trained network demonstrates robust performance in both visual and quantitative evaluation metrics.

Chapter 3

Proposed Method

This section describes our network. First, our network consists of two branches with different tasks and focuses on design. The attention guided context aggregation branch focuses on extracting low-level information(color, texture, etc.), while the transfer learning branch, focuses on extracting high-level information(object, event, etc.). We will look at the structural details(shown in Fig 3.1) of the two sub-networks and their design logic respectively, and analyze the benefits of each branch. Then we will introduce the loss functions used during the training process and explain their meanings respectively.

3.1 Network Architecture

As shown in Fig 3.1, our model consists of two branches: attention guided context aggregation(AGCA) branch and transfer learning branch. The dual-branch design allows each branch to process information separately to tackle different tasks with

same input. With the proper fusion method, this design has shown great performance on various deep learning task (Dai et al. 2020; Wu et al. 2020) when two branches are trained to be the complement of each other.

With this observation in mind, we design our dual-branch network with different purposes: the attention guided context aggregation branch aims to minimize the pixel loss during the conventional multi-scale process and the build-in attention mechanism will further guide the branch to increase the visibility of low-level features inside the heavy hazy area. The transfer learning branch, on the other hand, helps to extract robust feature maps from the input with pre-trained weight; the multi-scale encoder-decoder design is more capable of extracting high-level representation. The fusion tail is also properly tested, concatenating both high-level and low-level information from two branches, and fuse this complementary information and return the dehazed output.

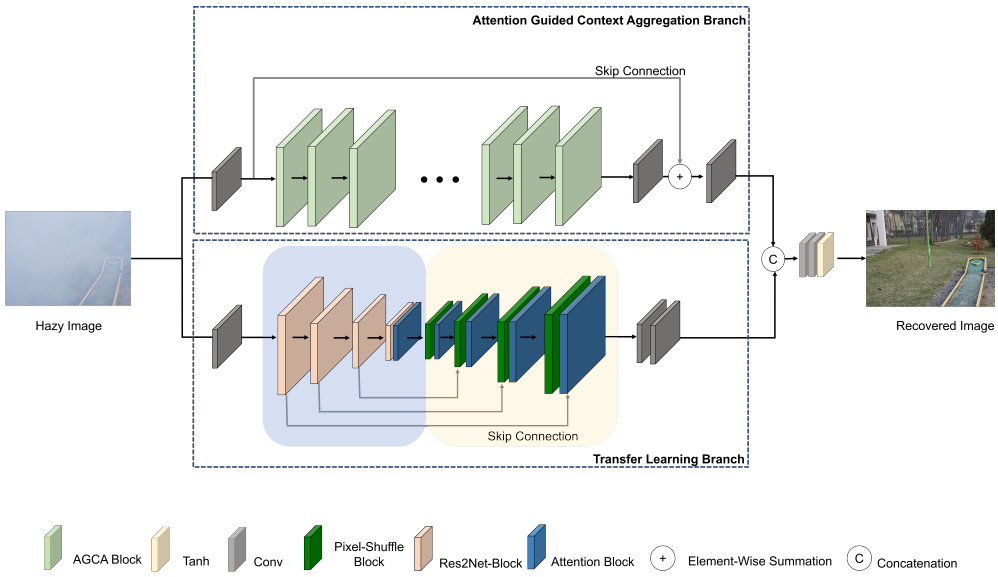

Figure 3.1: A representation of our model architecture. The model consists of two branches: attention guided context aggregation(AGCA) branch and transfer learning branch. AGCA branch consists of 14 AGCA blocks.

3.1.1 Attention Guided Context Aggregation Branch

In traditional CNN, to increase the receptive field of the network while decreasing the size of feature maps, sub-sampling operation is employed. As a result of the reduced size of feature maps, the number of feature maps can be increased without overloading the hardware memory, resulting in a bigger context for the final prediction through the network.

Due to the complex distribution of non-homogeneous haze, dehazing algorithms often suffer from loss of low-level details such as edge and corner or color distortion. We argue that these phenomena are closely related to pixel loss during the side-effect of handling the multi-size feature map using sub-sampling operation.

Inspired by the top path of (Zotti et al. 2017) where no sub-sampling operation is performed to keep the feature map sizes constant, we aim to design the attention guided context aggregation branch with a similar strategy within this branch to override the side effect of sub-sampling. However, without using the sub-sampling operation, the receptive field of the feature map is very small. Thus we adopt the dialed-convolution (Yu and Koltun 2015) within each AGCA layer(shown in Fig 3.2) to further increase the receptive field. Our final design of AGCA block is based on the SDCAB block (Deng et al. 2019) from a deraining task.

As shown in Fig 3.2, each AGCA block consists of two AGCA layers and skip connection, the skip connection further ensures the maximum information flow though the 14 AGCA blocks. The AGCA block consists of three scales of dilated convolution whose dilation scales are 1, 3, 5 respectively. We concatenate the output features right after to ensure that the most significant feature can be extracted.

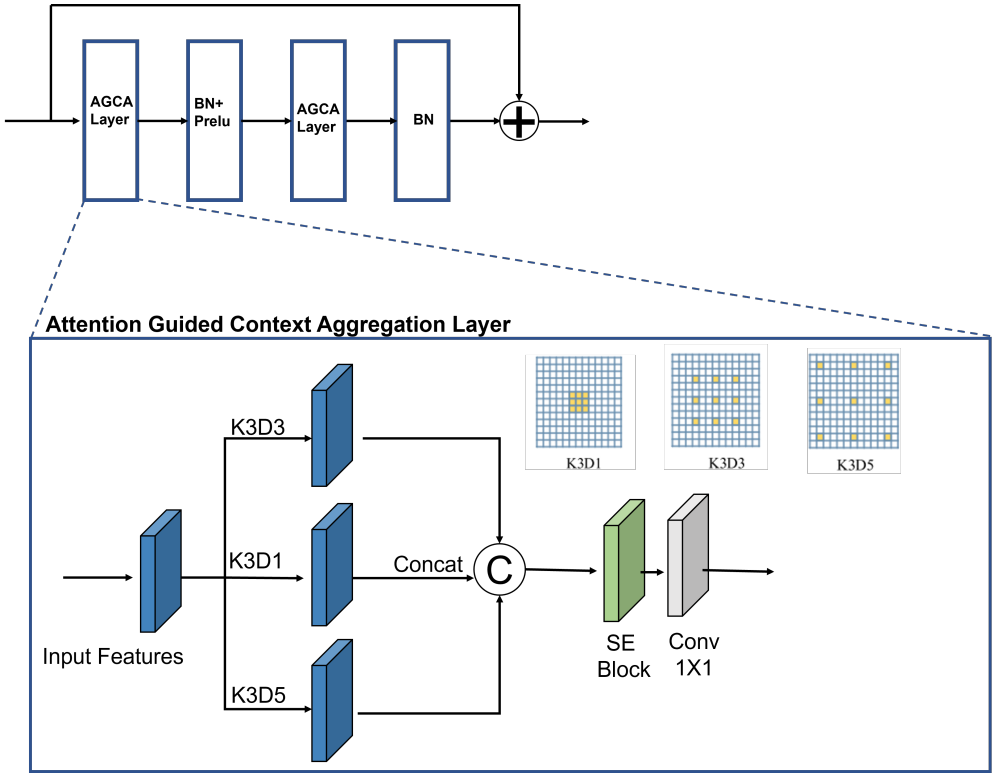

Figure 3.2: Detail structure of attention guided context aggregation block. Each block consists of two AGCA layers and a skip connection to ensure a large receptive field and maximum information flow. The illustration of detailed dilated operation has been shown on the right-hand-side of AGCA layer. (Please note that the channel attention block in green color will only be used at the first layer of the AGCA block due to computation cost.)

A simple channel attention (Hu et al. 2018) is used at the first layer of each AGCA block after the concatenation to highlight salient features. Finally, we utilize a 1 x 1 convolution to reduce the feature dimensions. Please note, the channel attention has only been added to the first layer of each AGCA to bring balance between computational cost and performance. In the end, a long skip connection has also been adopted between the first and last AGCA block to overcome the gradient vanishing problem (Hochreiter 1998).

Despite the effort to use feature maps at the same resolution and usage of dilated convolution, the AGCA branch still suffers from over-fitting problems during the training stage due to the small dataset provided during NTIRE challenge. To leverage the full power of AGCA branch and bring dynamic between high-level and low-level features, we are motivated to construct another branch that aims to transfer the knowledge learned from another larger dataset to extend learning ability of the whole system.

3.1.2 Transfer Learning Branch

Our transfer learning branch aims to help the AGCA branch overcome the over-fitting problem during the training stage by utilizing extra prior knowledge gained from the image classification task. To be more specific, we use Res2net (Gao et al. 2019) pretrained on ImageNet (Deng et al. 2009) as the encoder. The detail structure of Res2Net block can be seen from Fig 3.4.

Compared to the conventional bottleneck block, Res2Net block is known for its ability to express multi-scale features at the fine granularity level and increase the receptive field of each layer. This feature is consistent with our requirement for the non-homogeneous dehazing network to increase the receptive field for better visibility of the complex haze. As can be seen from Fig 3.4, after the input features pass the first 1x1 convolution layer, we divide the feature map to s equal subset, where s is a control parameter for scale dimension originally proposed by the (Gao et al. 2019); for our branch, $s = 4$. We define the subset as x_i, $i \in \{1, 2, \ldots, s\}$. Each feature subset has the same feature size but only a $1/s$ of the number of channels compared to the input features. Besides x_1, all other sub-features x_i

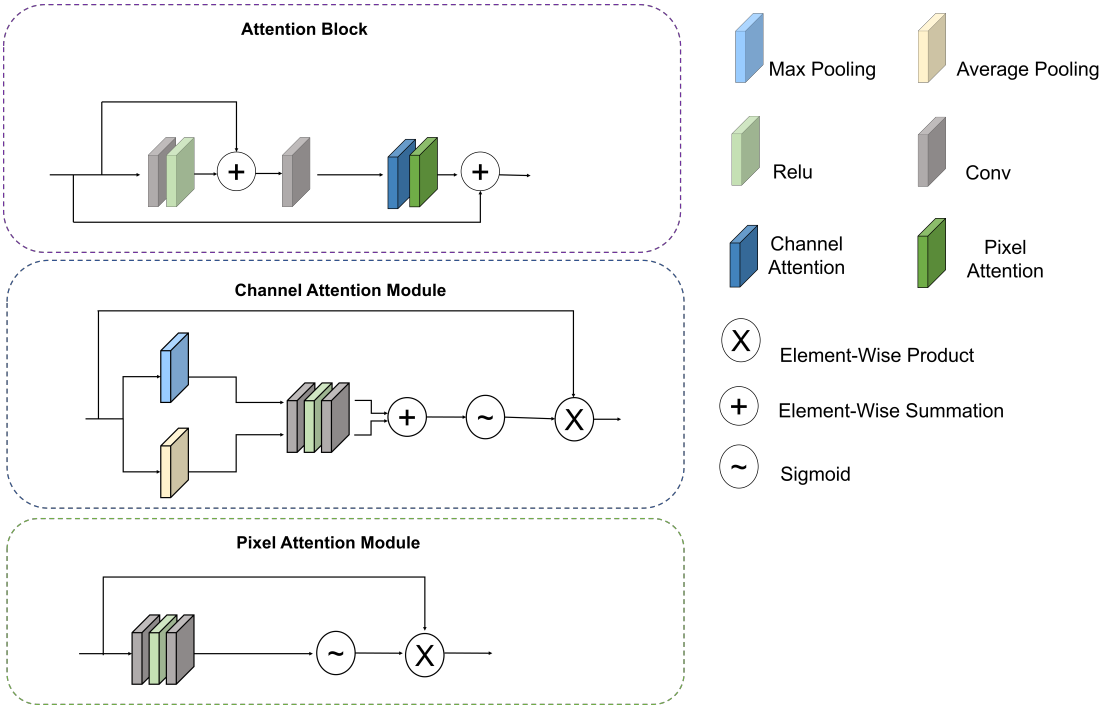

Figure 3.3: Attention Module

have corresponding 3x3 convolution kernel, denoted by $K_i()$. The output of K_i is denoted by y_i. Each 3x3 convolution operation can potentially accept all the feature information on its left, and each output can increase the receptive field, so each Res2Net output can obtain different number and combination of receptive field sizes/scales. Thus we can express y_i as:

$$y_i = \begin{cases} x_i & i = 1; \\ K_i\left(x_i\right) & i = 2; \\ K_i\left(x_i + y_{i-1}\right) & 2 < i \leq s. \end{cases} \tag{3.1}$$

As for the decoder part, we adopt PixelShuffle (Shi et al. 2016) layer to up-sample the feature map to the appropriate size. The main function of pixel shuffle

layer is to obtain a high-resolution feature map by convolution and multi-channel reorganization of the low-resolution feature map. This method was originally proposed to solve the problem of image super-resolution. It has become an effective way of up-sampling feature maps to original resolution.

Inspired by (Wu et al. 2020), we also adopt an attention module to further guide information extraction process during handling non-homogeneous haze. We have illustrated the attention module in Fig 3.3. The channel attention module is modified from the original channel attention module (Wu et al. 2020) by adding a max-pooling layer in parallel with average pooling layer. This design is based on CBAM channel attention (Woo et al. 2018), which shows that max-pooling layer can infer finer channel-wise attention on distinctive object features. For both channel and pixel attention, we utilize the sigmoid function as activation function. The feature first passes through the channel attention module then passes to the pixel attention. A skip connection is adopted to preserve more information.

3.1.3 Fusion Tail

The fusion tail takes the output feature maps generated by two distinct branches and outputs a clear image. To be more specific we have chosen two 7x7 convolution kernels followed by a tangent activation function to provide sufficient learning parameters.

3.1.4 Loss Function

To train the proposed network in a supervised manner, we employ the smooth L1 loss, MS-SSIM loss and perceptual loss.

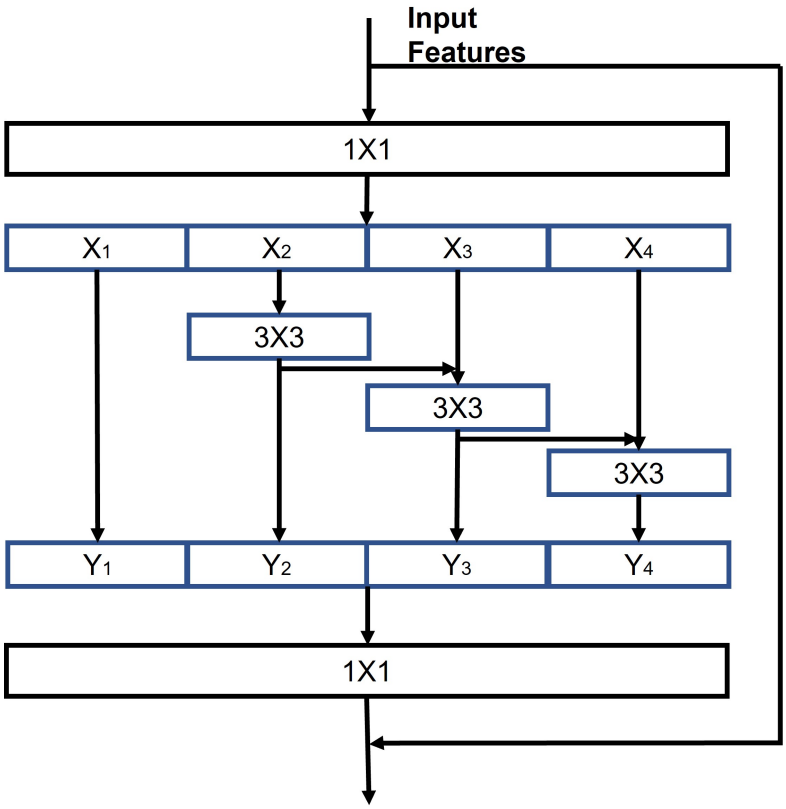

Figure 3.4: Res2Net Module

Smooth L1 aims to provide a quantitative measurement of error distance between the predicted image and ground truth image. It is a robust L1 norm that is less sensitive to outliers than the MSE since L1 can prevent gradient explosions (Girshick 2015). On the other hand, perceptual and MS-SSIM loss ensure that the visual quality of the predicted image is similar to the ground truth by providing high-level feature space and measurement of structural similarity respectively. The total loss of the proposed network is computed by a weighted sum of the aforementioned loss to enhance training performance.

Smooth L1 Loss. Smooth L1 loss can be considered as a combination of L1-loss (Zhu et al. 2015) and L2-loss (Zhao et al. 2016). It behaves like L1 loss

with small absolute errors and behaves like L2 Loss when the absolute value of the argument is close to zero. Smooth L1 loss has been widely used in various image restoration tasks (Deng et al. 2020; Wu et al. 2021), and it can be express as:

$$\mathcal{L}_{smooth-L1} = \frac{1}{N}\sum_{y=1}^{N}\sum_{i=1}^{3}\alpha\left(\hat{Y}_i(z) - Y_i(z)\right), \tag{3.2}$$

where $\hat{Y}_i(z)$ and $Y_i(z)$ represent the intensity of the i-th channel of each pixel z in the reconstructed haze-free image and in the ground truth, respectively. N indicates the number of pixels total, where α can be defined as follows:

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

Perceptual Loss. Besides the pixel-level loss, perceptual loss compares the features from the convolution of real pictures with those from the convolution of the generated picture, and brings high-level feature space (content and global structure) closer. We use VGG19 (Simonyan and Zisserman 2014) pre-trained weight on ImageNet (Deng et al. 2009) as a loss network, and we compute the feature losses at layers 2, 7, 12, 21 and 30 to ensure perceptual quality. The loss function can be described as:

$$\mathcal{L}_{\text{per}} = \sum_{j=1}^{3}\frac{1}{C_j H_j W_j}\left\|\phi_j(y) - \phi_j\left(y_t\right)\right\|, \tag{3.4}$$

where y and y_t are restored image and ground truth respectively. H_j , W_j , and C_j denote the height, the width, and the channel of the feature map in the j-th

layer, ϕ_j is the activation of the j-th layer.

SSIM loss. Furthermore, we employ a structural similarity loss (SSIM) (Wang et al. 2003) that is intended to reconstruct the RGB images through improving the structural similarity index. In contrast to predictions made without applying SSIM loss, the resulting images are more perceptually acceptable. And the loss function can be defined as :

$$L_{SSIM} = 1 - F_{SSIM}(\hat{Y}_i - Yi), \tag{3.5}$$

where F denotes the function of calculating structural similarity index.

Total loss. The total loss is described as:

$$\mathcal{L}_{total} = \lambda_{smooth-L1}\mathcal{L}_{smooth-L1} + \lambda_{SSIM}\mathcal{L}_{SSIM} + \lambda_{per}\mathcal{L}_{per}, \tag{3.6}$$

where $\lambda_{smooth-L1}$, λ_{SSIM}, λ_{per} are the hyper-parameters used to equally weigh different losses during the training stage.

Chapter 4

Experiments

In this section, we will first introduce dataset details, experiment setting, and evaluation metrics. Then, a series of ablation studies are conducted to demonstrate the benefit of each component in AGCA-Net. Finally, we compare our method with the other state-of-the-art dehazing algorithms. In the last part, AGCA-Net is tested extensively on synthetic and real-world datasets to demonstrate its effectiveness in dehazing quantitative results and qualitative effects.

4.1 Training and Testing Dataset

There are two categories in our training dataset: ASM-based synthetic dataset and real-world camera-based dataset. These datasets contain different haze distributions and have been widely used in various dehazing algorithms.

4.1.1 ASM-Based Dataset

Since there aren't many real-world hazy photos and their haze-free counterparts, data-driven dehazing methods often rely on synthetic hazy images, which can be derived by appropriately choosing the scattering coefficient and the atmospheric light of clear images generated using the ASM model.

RESIDE (Li et al. 2018a), which is an abbreviation of Realistic Single Image Dehazing, is a large-scale benchmark dataset consisting of both synthetic and real-world hazy images. RESIDE contains four subsets with diverse data sources and image content. For the ASM-based dataset, we adopt the widely used Indoor Training Set(ITS) and indoor Synthetic Objective Testing Set(SOTS) of RESIDE as our training and testing set respectively. The Indoor Training Set(ITS) of RESIDE contains 13990 hazy image pairs, generated from 1399 clear indoor images using atmospheric spherical scattering model with $\beta \in [0.6, 1.8]$ and $A \in [0.7, 1.0]$. The d(x) is derived from the (Silberman et al. 2012) and (Scharstein and Szeliski 2003). The Synthetic Objective Testing Set (SOTS) is utilized for testing, comprised of 500 hazy images pairs.

4.1.2 Real-World Dataset

Real-world datasets are generally hard to collect thus often appear in small scale, thus we try to utilize as much resources as we can to deal with the over-fitting problem when training with small scale dataset. The real-world dataset we adopt all utilize professional haze-generated machines to simulate the real-world haze. They can be grouped into two categories based on the distribution characteristics of the haze: homogeneous and non-homogeneous haze.

For the homogeneous haze real-world dataset, we adopt the Dense-Haze dataset (Ancuti et al. 2019a) generated for NTIRE2019 Dehazing Challenge(Ancuti et al. 2019b). DENSE-HAZE dataset contains 45 training pairs, 5 testing pairs and 5 validation pairs captured in a dense homogeneous haze environment. We use the 45 training pairs together with 5 additional validation pairs as the training set, and 5 testing pairs as the testing set.

For the nonhomogeneous haze real world dataset, we adopt the NH-Haze (Ancuti et al. 2020a) dataset obtained from NTIRE2020 (Ancuti et al. 2020b) Dehazing Challenge and NH-Haze2 dataset generated for NTIRE2021 (Ancuti et al. 2021) Dehazing Challenge. NH-Haze contains 45 training pairs, 5 validation pairs and 5 additional testing pairs. NH-HAZE2, on the other hand, only contains 25 nong-homogeneous hazy images and the corresponding ground truth image with no additional validation and testing pair. Thus, we choose 20 of the 25 pairs as the training set, and the remaining 5 pairs as the testing set.

4.2 Evaluation Metrics

Every method in this section is trained in the same way with AGCA-Net for fair comparison, and then evaluated on the test sets mentioned above. For the quantitative comparison, two objective quality metrics are used: Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Assuming that a dehazed image and the corresponding ground-truth are given, both PSNR and SSIM are calculated to determine how similar an image is to its reference both pixel-wise and structurally.

Mathematically, for an $m \times n$ dehazed image K and its ground truth image I, the PSNR is defined as:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right), \tag{4.1}$$

where MSE is described as,

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2. \tag{4.2}$$

SSIM can be defined as:

$$SSIM(I,K) = \frac{(2\mu_I\mu_K + c_1)(2\sigma_{IK} + c_2)}{(\mu_I^2 + \mu_K^2 + c_1)(\sigma_I^2 + \sigma_K^2 + c_2)}, \tag{4.3}$$

where μ_I, σ_I^2 are the mean value and variance of I, μ_K, σ_K^2 are the mean value and variance of K, σ_{IK} is the covariance of I and K. $c_1 = (a_1 L)^2, c_2 = (a_2 L)^2$ are two constants with a_1=0.01 and a_2=0.03 being default values respectively and L represents the range of pixel value.

4.3 Training Details

We adopt the same training strategy regardless of the differences in the characteristics of each dataset. First, We use simple data augmentation strategy to increase training samples to 8 times of original image counts. In order to do this, we crop 256x256 size patches and provide random rotation of 90, 180, 270 degrees along with horizontal or vertical flip. Second, for best optimization result we utilize Adam optimizer (Kingma and Ba 2014) with initial learning rate 0.0001, and

$\beta_1 = 0.9$ and $\beta_1 = 0.999$. The batch size is set to be 4 and we conducted our all of experiments on a single Nvidia V100 GPU. Our method is implemented using Pytorch library. Finally, to balance the loss in a smoother loss surface, we set $\lambda_{smooth-L1} = 1$, $\lambda_{SSIM} = 0.22$, $\lambda_{per} = 0.36$ respectively.

4.4 Ablation Study

In this study, we consider different configurations and module combinations of the attention guided context aggregation network(AGCA-Net). During the study, a combination of three factors goes into ablation study. First, we test the effectiveness of ImageNet pre-trained weights. Second, we test the effectiveness of the AGCA-block. Finally, we test the effectiveness of combined the dual branch together. The ablation configuration is shown below:

1. TLB without pre-trained weight: only use the randomly initialized transfer learning branch.

2. TLB with pre-trained weight: only use the transfer learning branch, but loaded with ImageNet pre-trained weight

3. AGCA-10: only use the AGCA branch constructed by 10 attention guided context aggregation blocks.

4. AGCA-12: only use the AGCA branch constructed by 12 attention guided context aggregation blocks.

5. AGCA branch: only use the AGCA branch proposed in this paper, with 14 attention guided context aggregation blocks.

6. TLB without pre-trained + AGCA branch: use both transfer learning branch and proposed AGCA branch without loading ImageNet pre-trained weight.

7. TLB + AGCA branch: use both pre-trained transfer learning branch and AGCA branch.

Methods	Pre-trained	PSNR	SSIM
(1)TLB	-	18.01	0.553
(2)TLB	$\surd$	19.75	0.783
(3)AGCA-10	-	19.12	0.694
(4)AGCA-12	-	19.97	0.736
(5)AGCA	-	20.05	0.819
(6)TLB+AGCA	-	20.15	0.862
(7)TLB+AGCA	$\surd$	20.49	0.879

Table 4.1: Quantitative comparison of ablation study. TLB represents the transfer learning branch, AGCA represents the 14-layers attention guided context aggregation Branch. " $\surd$ " represents the branch loaded with ImageNet pre-trained weight, " - " represents that no pre-trained weight was loaded.

Please refer to Fig 4.1 for visual comparison result of the NH-HAZE2 from NTIRE2021. We conduct our experiment on the test set of NH-HAZE2. The quantitative results can be found in Table 4.1. From Table 4.1, we can clearly see the improvement in both PSNR and SSIM when using pre-trained ImageNet weight (by comparing (1) and (2)), thus suggesting that utilizing knowledge transfer can indeed boost dehazing performance. On the other hand, by comparing (3),(4),(5), we also can observe that both PSNR and SSIM increase with more AGCA blocks. The AGCA branch (5) proposed in this paper with 14 AGCA blocks, achieves the maximums of 20.05 dB and 0.819 in terms of PSNR and SSIM, which indicates the effectiveness of the AGCA blocks even without using ensemble strategy.

(6) and (7) are dual-branch model designs, and both have superior dehazing performance compared to single branch designs. By comparing (6) and (7), we find by loading the network with pre-trained ImageNet weight, the dehazing performance is boosted 0.34 dB in terms of PSNR, which shows the strong predictive ability of the pre-trained transfer learning branch. With our full model, we achieve 20.49 dB and 0.879 in PSNR and SSIM. Both quantitative and visual results indicate the effectiveness of our design.

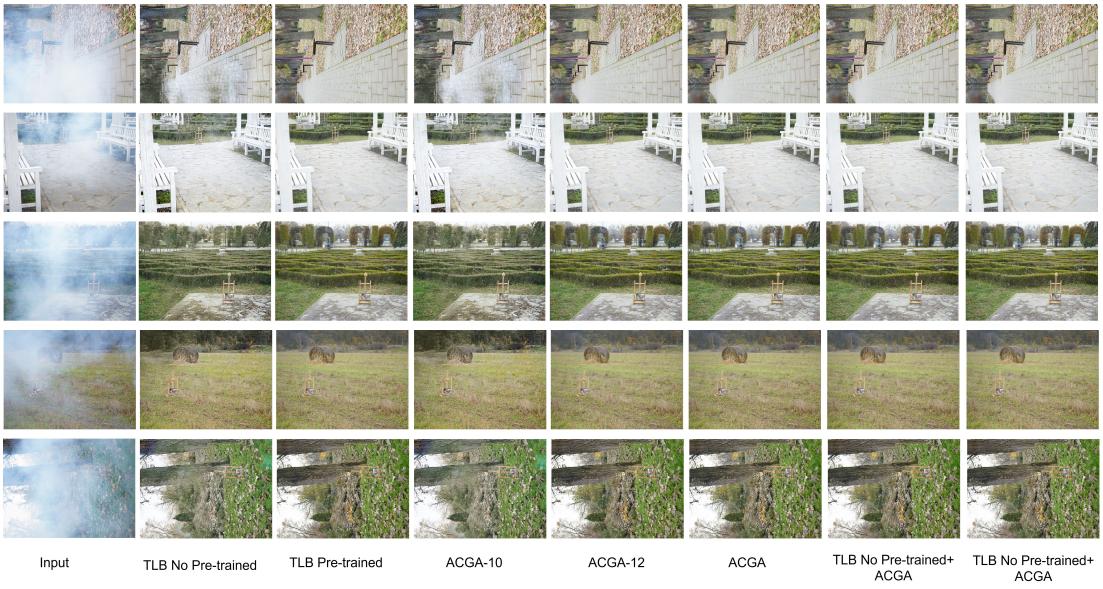

Figure 4.1: The visual comparison of the NTIRE2021 NH-HAZE2 ablation study

4.5 Comparisons with the State-of-the-art

We compare our proposed method with state-of-art methods on both synthetic ASM based dataset (Li et al. 2018a) and real-word camera based dataset (Dense-Haze (Ancuti et al. 2019b), NH-Haze (Ancuti et al. 2020b), and NH-Haze2 (Ancuti et al. 2021)). Those SOTA methods include: DCP (He et al. 2010), AOD-NET (Li et al. 2017), GCA-Net (Chen et al. 2019), FFA-net (Qin et al. 2020) and KTDN

(Wu et al. 2020). We train DCP, AOD-Net and KTDN on RESIDE dataset, and obtain testing results of GCA-Net and FFA-Net from release paper and corresponding testing code. As for the real-world dataset, all methods are substantially trained to gain best result. Although our method does not significantly surpass the SOTA methods, the competitive results prove the effectiveness of our method. Both visual and quantitative results of mentioned methods can be found in the following section.

4.5.1 Result on RESIDE Dataset

Qualitative Visual Result Fig 4.2 shows the qualitative visual comparison of the indoor images from RESIDE. DCP tends to cause severe color distortion (darker than ground truth) due to the inaccurate estimation of ASM model, which harms the output quality of the image. While the AOD-Net does largely overcome the color distortion problem, it leaves much of the hazy effect un-removed. GCA-Net surpasses the above two methods and removes most of the haze, but there are still region suffering from color distortion. KTDN and our method have better visual results by removing most of the haze without losing important image features and causing color distortion. The FFA-net has the best visual result among all methods.

Quantitative Result Table 4.2 shows the quantitative comparison of all the methods. DCP and AOD-net with 18.93 dB and 19.06 dB PSNR respectively, have the worst visual quality as well. All other methods have PSNR exceeding 30 dB. FFA achieves the best performance on PSNR, surpassing our method by 3.07 dB. Our method ranked second place in the testing.

Methods	DCP	AOD-Net	GCANet	FFA	KTDN	Ours
PSNR	18.93	19.06	30.15	**36.69**	30.23	<u>33.62</u>
SSIM	0.882	0.852	0.975	**0.988**	<u>0.982</u>	0.936

Table 4.2: Quantitative comparisons of RESIDE(ITS) dataset. The best results are in **bold**, and the second best are with <u>underline</u>.

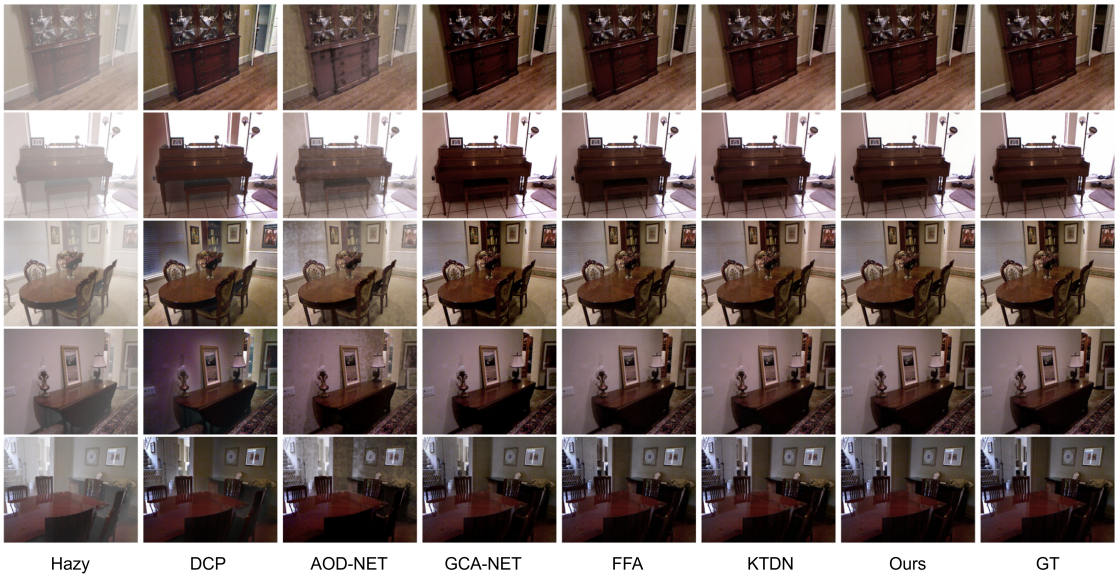

Figure 4.2: Qualitative visual evaluation on RESIDE(ITS).

4.5.2 Result on Dense-Haze Dataset

Qualitative Visual Effect Result From Fig 4.3, we clearly see that due to the dense haze of input images, there is a serious loss of details, textures, edges, and colors in the input image. As the result of increasing dehazing difficulty, there are color distortions in DCP such that the colors of results are bluer than a true sense, and the priors are clearly violated in the certain area. AOD-Net generated output is darker than ground truth, and large portion of the dense haze also remains un-removed, and most of details of the image are still bury under haze. GCA-Net also has difficulty in removing the dense haze especially in areas with a similar color. KTDN and our method can remove most of the haze, but when haze is present in

Methods	DCP	AOD-Net	GCANet	FFA	KTDN	Ours
PSNR	11.15	13.01	12.35	**16.26**	15.25	<u>15.95</u>
SSIM	0.423	0.471	0.469	0.526	<u>0.531</u>	**0.552**

Table 4.3: Quantitative comparisons of Dense-Haze dataset. The best results are in **bold**, and the second best are with <u>underline</u>.

high concentrations, they may not be able to remove it completely. FFA has the best visual result among all methods, but we can still observe severe loss of details and color distortion in the most concentrated hazy area.

Quantitative Result As shown in Table 4.3, our method ranked 2nd place in terms of PSNR and 1st place in SSIM, with 15.95 dB and 0.552 respectively. FFA-net has best PSNR of 16.26 dB and the best visual performance overall. DCP, AOD-net and GCA-net, which produced non-visually pleasing results, also shown struggled performance in both PSNR and SSIM.

4.5.3 Result on NH-Haze Dataset

Qualitative Visual Result From Fig 4.4, we can see clearly that much of the color information has been lost on DOP, which causes DCP to produce a bluer color. Also, we can obverse remaining haze and artifacts. AOD-net cannot remove the haze effectively as a big portion of area is still covered by non-homogeneous haze. GCA-net also suffer from lost of details and shows deviate color. In addition, due to the huge GPU usage, we have to split the input image to 4 smaller chunks and combine them together after individual inference, resulting in more color distortion. Although FFA shows good results as it has removed a large portion of the haze, it also suffers from GPU overload problem, but relying on the strong dehazing ability, the FFA experiences a lighter "checkerboard" effect compared to

Methods	DCP	AOD-Net	GCANet	FFA	KTDN	Ours
PSNR	12.92	12.94	17.29	18.79	<u>20.31</u>	**20.45**
SSIM	0.485	0.395	0.601	0.638	**0.736**	<u>0.737</u>

Table 4.4: Quantitative comparisons of NH-HAZE 2020 dataset. The best results are in **bold**, and the second best are with <u>underline</u>.

AOD-net. KTDN can remove most of the complex haze, and reconstruct most of the sharp edges with a little artifact. Our method produces similar visual result as KTDN by preserving most of the details but also has a more vivid color. This again shows the effectiveness of our design by restoring both high-level and low-level information.

Quantitative Result DCP and AOD-Net have the worst performance both visually and quantitatively, with 12.92 dB and 12.94 dB in PSNR respectively. Our method achieves the best PSNR score of 20.45 dB, 0.14dB more than the KTDN. The results can be found in Table 4.4.

4.5.4 Result on NH-Haze2 Dataset

Qualitative Visual Effect Result From Fig 4.5, DCP still produces a bluish image and artifact. AOD-Net performs better than NH-Haze dataset by removing more haze, but still can't produce a visually pleasing result. Both FFA and KTDN manage to remove most of the haze but still suffer color distortion in the dense haze area. Our method achieves the best visual results by removing most of the haze and minimum color deviation among all methods.

Quantitative Result Our method also achieves the best quantitative result in the dataset, with the highest PSNR score of 20.49 dB and SSIM 0.879. FFA ranked

Methods	DCP	AOD-Net	GCANet	FFA	KTDN	Ours
PSNR	12.56	13.98	19.01	<u>20.42</u>	20.38	**20.49**
SSIM	0.405	0.5589	0.801	<u>0.858</u>	0.752	**0.879**

Table 4.5: Quantitative comparisons of NH-HAZE 2021 dataset. The best results are in **bold**, and the second best are with <u>underline</u>.

Methods	**Parameters**
AOD	1.7K
FFA	3.7M
GCA	0.7M
KTDN	40M
Our	68M

Table 4.6: Total number of parameter of each method.

2nd place with 1.63 dB gained compared to NH-Haze. GCA-net also improves significantly compared to NH-Haze, with 19.01 dB. The results can be found in Table 4.5.

4.5.5 Runtime Comparison

We conduct this experiment on a single Nvidia V100 GPU, and show the result in Fig 4.6. We also provide the total number of parameters of each method in Table 4.6. Our proposed method takes about 0.382 seconds to dehaze an image with size 1200 x 1600, which is very close to KTDN with 0.35s. Also our method has 68M total parameters and it is the largest model among all other methods. Please refer to Table 4.6 for details.

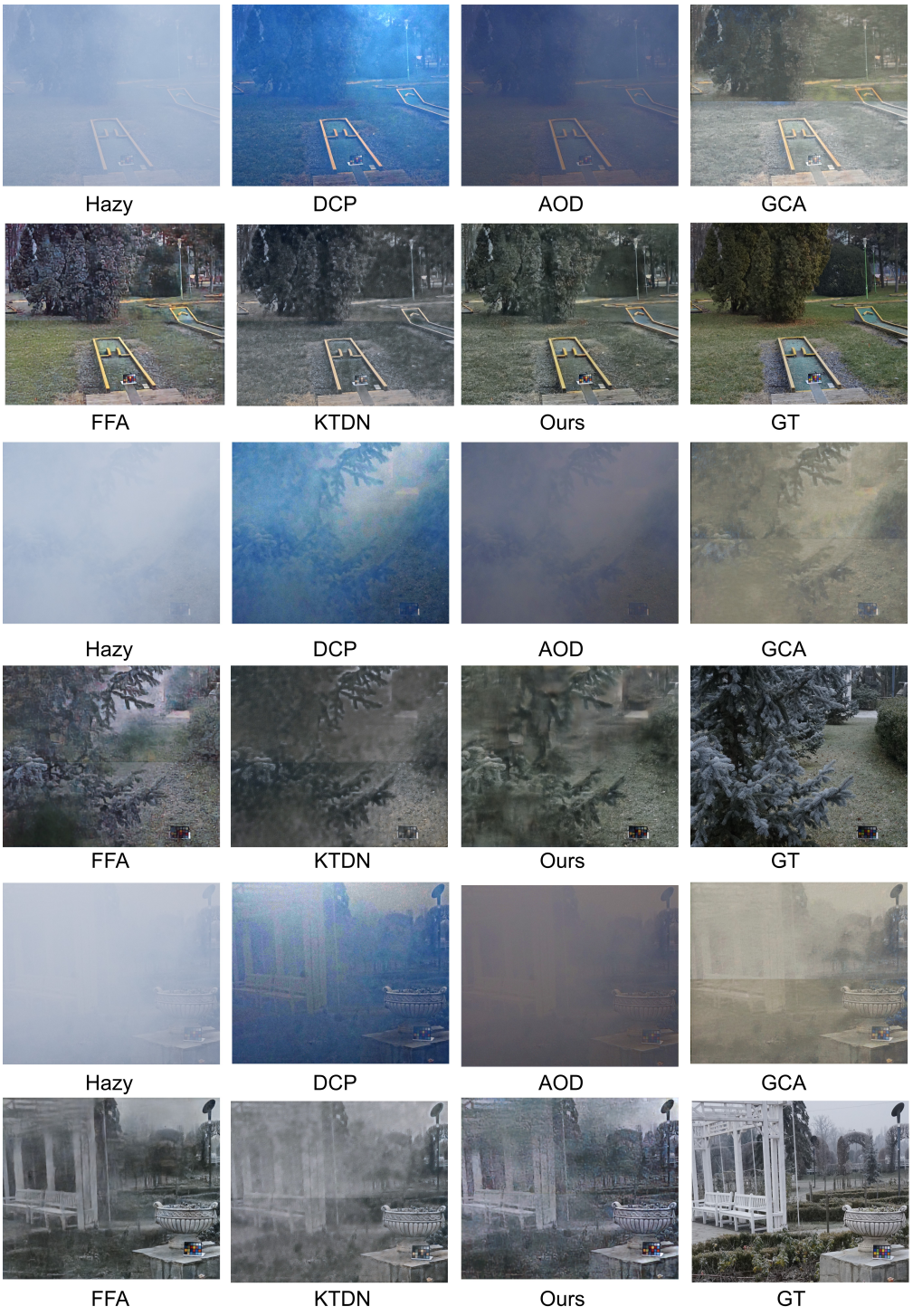

Figure 4.3: Qualitative visual evaluation on Dense-Haze.

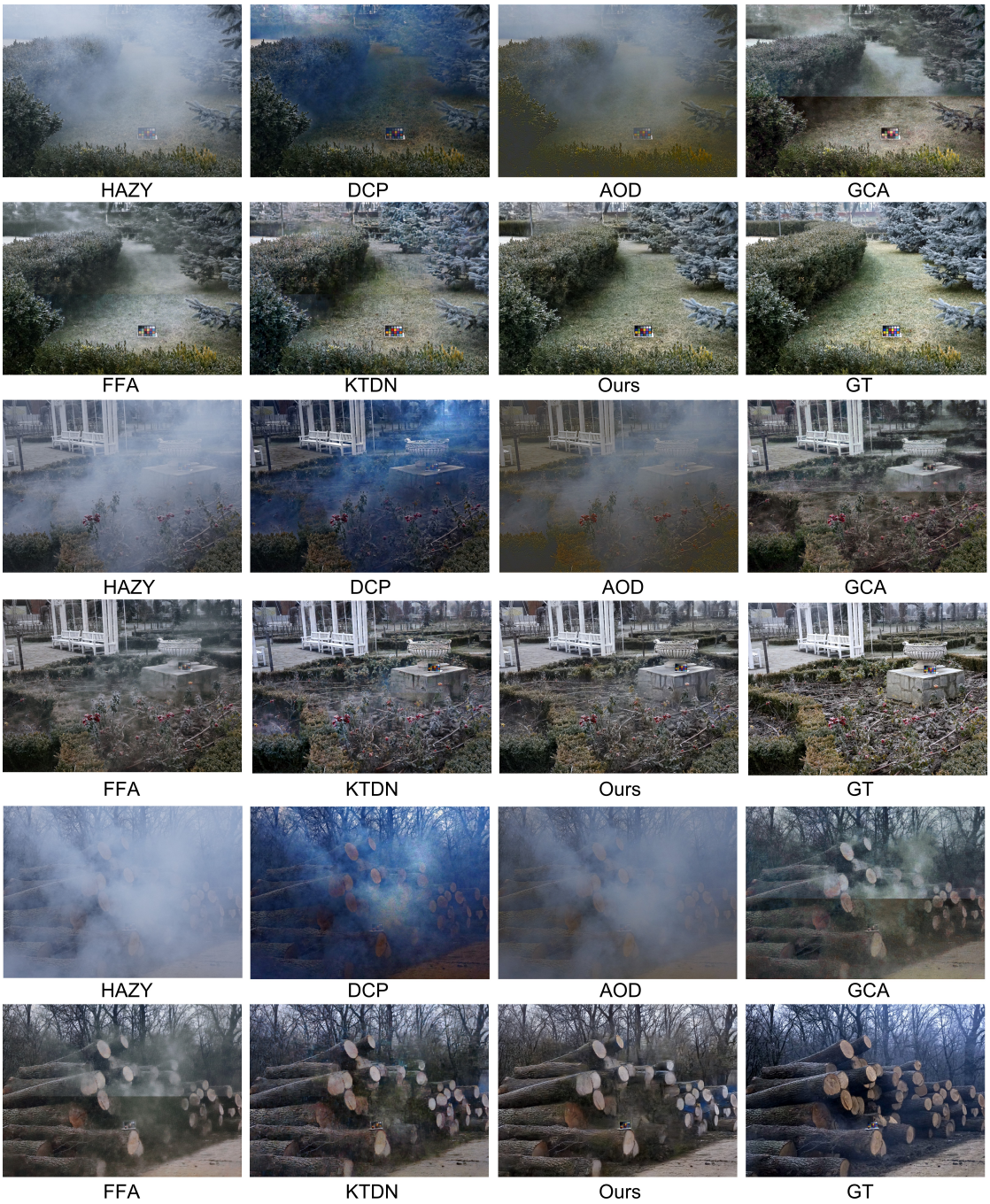

Figure 4.4: Qualitative visual evaluation on NH-Haze.

Figure 4.5: Qualitative visual evaluation on NH-HAZE2.

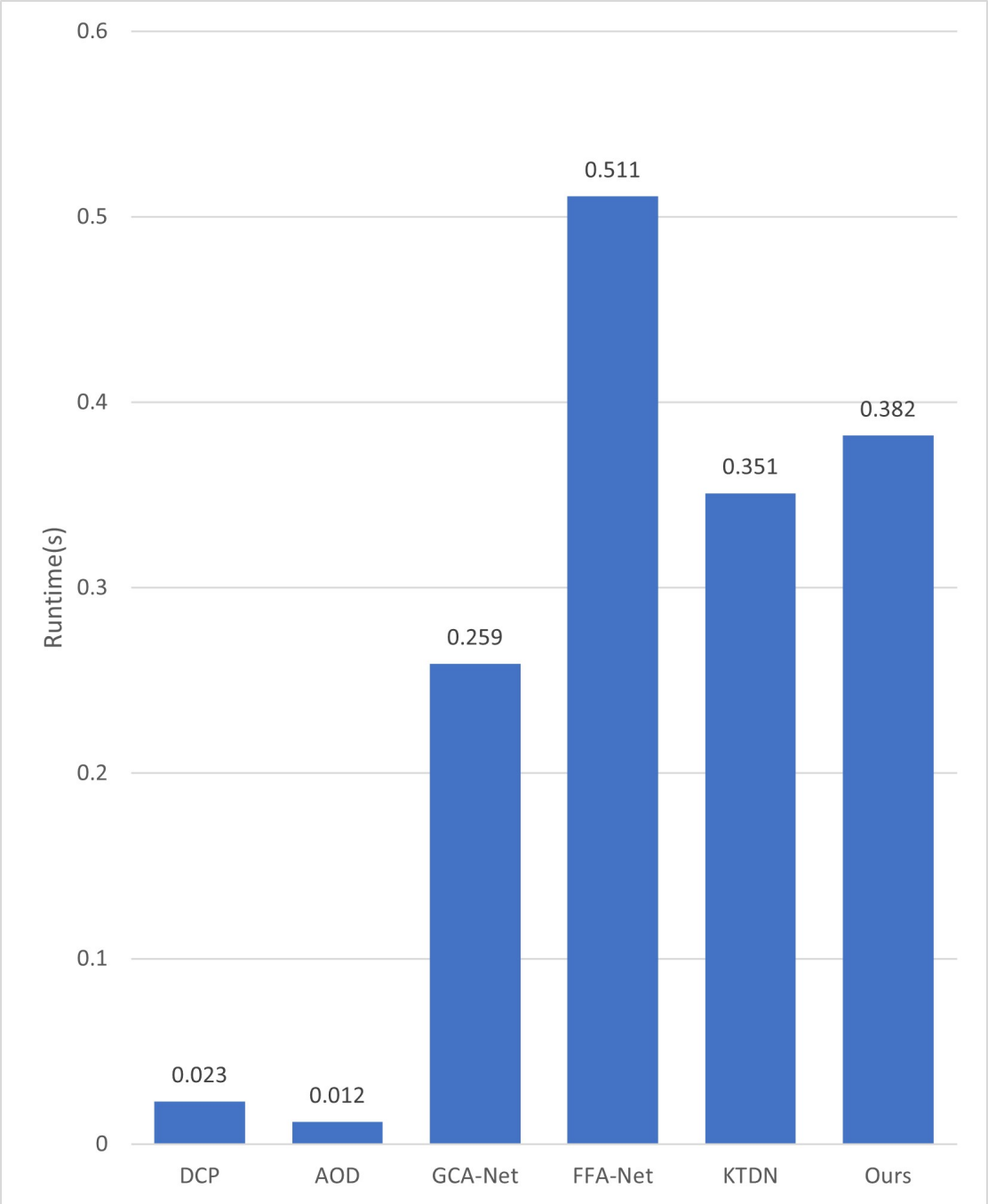

Figure 4.6: Compare the runtime performance of DCP, AOD, GCA, FFA, KTDN and our methods on NH-HAZE2.

Chapter 5

Conclusion

In this thesis, we have proposed a dual-branch attention guided context aggregation network, namely AGCA-Net. Due to the dual-branch design, we can focus on high-level and low-level information separately, then combine and balance the information between the two in the output dehazed image. Also, we utilize a transfer learning strategy in the transfer learning branch with an ImageNet pre-trained weight to remove haze even when there is only a small dataset available. Although our final result still falls short of the performance of state-of-the-art systems, a large number of experiments have confirmed our design to be feasible and effective.

Bibliography

Ancuti, C. O., Ancuti, C., Sbert, M., and Timofte, R. (2019a). Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In: *2019 IEEE international conference on image processing (ICIP)*. IEEE, 1014–1018.

Ancuti, C. O., Ancuti, C., and Timofte, R. (2020a). NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 444–445.

Ancuti, C. O., Ancuti, C., Timofte, R., Van Gool, L., Zhang, L., and Yang, M.-H. (2019b). Ntire 2019 image dehazing challenge report. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Ancuti, C. O., Ancuti, C., Vasluianu, F.-A., and Timofte, R. (2020b). Ntire 2020 challenge on nonhomogeneous dehazing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 490–491.

Ancuti, C. O., Ancuti, C., Vasluianu, F.-A., and Timofte, R. (2021). NTIRE 2021 nonhomogeneous dehazing challenge report. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 627–646.

Anvari, Z. and Athitsos, V. (2020). Dehaze-GLCGAN: unpaired single image dehazing via adversarial training. *arXiv preprint arXiv:2008.06632*.

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

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* 25(11), 5187–5198.

Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., and Hua, G. (2019). Gated context aggregation network for image dehazing and deraining. In: *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1375–1383.

Dai, L., Liu, X., Li, C., and Chen, J. (2020). Awnet: Attentive wavelet network for image isp. In: *European Conference on Computer Vision*. Springer, 185–201.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

Deng, Q., Huang, Z., Tsai, C.-C., and Lin, C.-W. (2020). Hardgan: A haze-aware representation distillation gan for single image dehazing. In: *European Conference on Computer Vision*. Springer, 722–738.

Deng, S., Wei, M., Wang, J., Liang, L., Xie, H., and Wang, M. (2019). DRD-Net: Detail-recovery Image Deraining via Context Aggregation Networks. *arXiv preprint arXiv:1908.10267*.

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., and Yang, M.-H. (2020). Multi-scale boosted dehazing network with dense feature fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2157–2167.

Du, Y. and Li, X. (2018). Perceptually optimized generative adversarial network for single image dehazing. *arXiv preprint arXiv:1805.01084*.

Engin, D., Genç, A., and Kemal Ekenel, H. (2018). Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 825–833.

Fu, M., Liu, H., Yu, Y., Chen, J., and Wang, K. (2021). DW-GAN: A Discrete Wavelet Transform GAN for NonHomogeneous Dehazing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 203–212.

Gao, S., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. H. (2019). Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*.

Girshick, R. (2015). Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K., Sun, J., and Tang, X. (Aug. 2010). Single Image Haze Removal Using Dark Channel Prior. *IEEE transactions on pattern analysis and machine intelligence* 33.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6(02), 107–116.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Judd, C. H. (1927). Generalized experience.

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

Li, B., Peng, X., Wang, Z., Xu, J., and Feng, D. (2017). An all-in-one network for dehazing and beyond. *arXiv preprint arXiv:1707.06543*.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., and Wang, Z. (2018a). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing* 28(1), 492–505.

Li, L., Dong, Y., Ren, W., Pan, J., Gao, C., Sang, N., and Yang, M.-H. (2019). Semi-supervised image dehazing. *IEEE Transactions on Image Processing* 29, 2766–2779.

Li, R., Pan, J., Li, Z., and Tang, J. (2018b). Single image dehazing via conditional generative adversarial network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8202–8211.

Liu, C., Fan, J., and Yin, G. (2020). Efficient Unpaired Image Dehazing with Cyclic Perceptual-Depth Supervision. *arXiv preprint arXiv:2007.05220*.

Liu, H., Wang, C., and Chen, J. (2021). Indirect Domain Shift for Single Image Dehazing. *arXiv preprint arXiv:2102.03268*.

Liu, X., Ma, Y., Shi, Z., and Chen, J. (2019). Griddehazenet: Attention-based multi-scale network for image dehazing. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7314–7323.

McCartney, E. J. (1976). Optics of the atmosphere: scattering by molecules and particles. *New York*.

Pei, Y., Huang, Y., and Zhang, X. (2019). Classification-driven Single Image Dehazing. *arXiv preprint arXiv:1911.09389*.

Qin, X., Wang, Z., Bai, Y., Xie, X., and Jia, H. (2020). FFA-Net: Feature fusion attention network for single image dehazing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, 11908–11915.

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*. Springer, 154–169.

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: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3253–3261.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

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.* Vol. 1. IEEE, I–I.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Shu, Q., Wu, C., Xiao, Z., and Liu, R. W. (2019). Variational regularized transmission refinement for image dehazing. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2781–2785.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In: *European conference on computer vision*. Springer, 746–760.

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

Tan, R. T. (2008). Visibility in bad weather from a single image. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee, 1398–1402.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, H., Liu, J., Xie, Y., Qu, Y., and Ma, L. (2020). Knowledge transfer dehazing network for nonhomogeneous dehazing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 478–479.

Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., and Ma, L. (2021). Contrastive Learning for Compact Single Image Dehazing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10551–10560.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhang, H. and Patel, V. M. (2018). Densely connected pyramid dehazing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3194–3203.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging* 3(1), 47–57.

Zhu, Q., Mai, J., and Shao, L. (2014). Single Image Dehazing Using Color Attenuation Prior. In: *BMVC*. Citeseer.

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

Zotti, C., Luo, Z., Humbert, O., Lalande, A., and Jodoin, P.-M. (2017). GridNet with automatic shape prior registration for automatic MRI cardiac segmentation. In: *International workshop on statistical atlases and computational models of the heart*. Springer, 73–81.