RESEARCH ARTICLE

Enhanced Multiscale Attention Network for Single Image Dehazing

WATARU IMAI, (Member, IEEE), AND MASAAKI IKEHARA, (Senior Member, IEEE)
Department of Electronics and Electrical Engineering, Keio University, Yokohama 223-8522, Japan
Corresponding author: Wataru Imai (imai@tkhm.elec.keio.ac.jp)

ABSTRACT Under severe weather conditions, the quality of the images taken outside is directly affected by floating atmospheric particles. To keep the quality of the images, haze removal methods play a critical role. The most difficult part of haze removal is removing the haze that spreads over the entire image. Many CNN-based methods have been proposed to remove the haze, and can be divided into two types. One is to use a multi-scale structure and the other is to stack layers. The former causes image degradation due to the loss of some of the original information in an image and the latter increases computational complexity due to not reducing the resolution. In addition, a large number of parameters is required to secure the expressive power of the model, which leads to a huge amount of memory. To tackle these problems, we tried to 1) downsample the image while saving parameters and maintaining the quality of the generated image, and 2) consider the information in the entire image to remove the haze. For the first problem, we tried to solve this by using a feature extractor that has been used in other tasks, learning to optimize the output image in low-resolution, and preparing kernels with various dilation rates to expand the receptive fields. For the second problem, we use the attention structure to determine which part of the image features should be focused on from the entire feature map. By incorporating such modules, our method achieves better results on both synthetic and real-world images when compared with state-of-the-art methods.

INDEX TERMS Attention, deep learning, multi-scale, single image dehazing.

I. INTRODUCTION

In recent years, the demand for high quality images has been increasing due to the spread of social networking services and the development of high-level computer vision tasks, such as segmentation [1], [2] and object detection [3], [4]. However, images available in the real world have many degradation factors such as weather conditions, noise caused by light and cameras, and blur. Haze is one of these degradation factors and single image dehazing is a research to remove the haze from an image. Haze is a phenomenon in which light is diffused by haze or smoke, and affects the contrast and color of the image. The physical haze model [5], [6], [7] is expressed by the following equation.

\[ I(x) = t(x)J(x) + (1 - t(x))A \]  

where, \( I \), \( J \), \( t \), and \( A \) are the input haze image, the clean image, the transmission coefficient, and the ambient light, respectively. Recently, due to the development of hardware, deep learning methods are widely used for single image dehazing and there are mainly two types of methods. One is to directly output a clean image in an end-to-end manner [8], [9], [10], [11], [12], [13], and the other [14], [15], [16], [17] is to output a haze free image by estimating the transmission coefficient and ambient light through a network and substituting these values into Eq. (1). Unlike other degradation factors, haze is greatly affected by distance and it is very difficult to train the network without considering the distance information. However, the depth information is required to obtain the correct label of the transmission map, which is...
not easy to obtain and almost impossible for outdoor images. Therefore, recently, end-to-end methods have become the mainstream. Nowadays, there are many high-performance end-to-end methods, but they have the following problems.

A. LIMITATION OF RECEPTIVE FIELD
Since haze spreads over the entire image, a large receptive field is essential for capturing this feature. In haze removal, CNN-based methods have achieved a significant improvement in quality, but they have difficulties in effectively extracting haze features because the convolution reconstructs the image by only considering local regions. Actually, in high resolution and dense haze datasets such as NH-HAZE [18] and Dense-HAZE [19], many methods cannot completely remove the haze. An image including haze is often not recognized as haze, or even if it is recognized and removed, haze remains in spots due to localization caused by the narrow receptive field. Recently, as the resolution of images is higher, it is very important to deal with this problem.

B. IMAGE DEGRADATION AND COLOR DISTORTION
To cope with the limitation of receptive field, most conventional methods [8], [9], [12] employ an encoder-decoder structure such as U-Net [20], tries to expand the receptive field by downsampling. However, downsampling causes the loss of pixel information, so it is necessary to recover the missing information in addition to the haze removal. Even if the haze is removed using a wide receptive field, the problem of restoring the background color still remains. Since there is almost no information on the color of the background objects in the input image, a high level feature extraction of the objects is necessary to restore their color. Therefore, while some methods are able to remove haze, few are able to accurately output the correct color of the object.

C. HEAVY PARAMETERS AND PROCESSING
In order to reduce image degradation and color distortion, networks have gotten wider and deeper. A wide network increases the number of filters to ensure the representativeness of the model, while a deep network increases the number of layers. The former requires a huge number of parameters, and the latter causes the problem of increased computational complexity. MSBDN [12], a wide network, proposed a dense feature fusion module using back-projection, which is often used in super-resolution [21], [22], [23]. Although this module improves the quantitative evaluation while downsampling the image, the number of parameters for this network, which processes many feature maps densely, has increased to 29M. On the other hand, FFA-Net [10] is one of the deep networks and tries to deal with the problem of image degradation by processing the image without downsampling. However, since the image is processed in the original resolution, the amount of computation increases further, which leads to memory overload and processing speed reduction.

To address these issues, we propose Enhanced Multi-Scale Attention Network (EMSAN) while reducing the parameters. Furthermore, various modules, such as attention, are added to achieve more global and advanced feature extraction and processing.

To expand the receptive field, it is necessary to reduce the size of the image and process it, but information included in the image is lost due to downsampling during encoding. To achieve this trade-off, we propose a Mixed Encoder(ME). ME is an encoder consisting of a pre-trained VGG16 and Dense Net. By using these networks as the encoder, the ME not only avoids missing information during encoding, but also performs high-level feature extraction and prevents color distortion. Additionally, we propose the Multi Output Branch(MOB) structure to improve the accuracy of low-scale branches in a multi-scale network, which can reduce the deterioration. MOB generates low resolution output images from the low-scale features through a Refine Block [8]. By taking the loss between those output images and the downsampled ground truth, the feature extraction at lower scales is optimized.

In addition, we have incorporated a new block structure, the Enhanced Feature Attention module inspired from the FA module proposed in FFA-Net [10]. This module uses dilated convolution, pixel attention, and channel attention to consider the entire feature map. The previous vertical pixel and channel attention structures are likely to cause bottlenecks in one of them, so we address this problem by parallelizing them.

Processing at the original resolution has a significant impact on the final output. Therefore, instead of using a structure based on the convolutional layers, which consider local regions, we develop a structure based on attention so that the information of the entire image can be considered. The attention part is structured around the Multi-head Self Channel Attention(MH-SCA) block. The conversion of the Fully Connected layer of Channel attention to Multi-head Self Attention allows the processing according to the semantic features of the image.

We summarize the contributions of our work as follows.

- We propose Enhanced Multi-Scale Attention Network (EMSAN) for single image dehazing. EMSAN achieves the highest quantitative and qualitative evaluation on several datasets when compared with conventional methods while saving the number of parameters.

- We introduce the Mixed Encoder(ME) and Multi Output Branch(MOB). These modules optimize advanced feature extraction and low-scale networks, and reduces information degradation due to downsampling and color distortion.

- A new block structure, Enhanced Feature Attention module, is proposed to capture more global features. This incorporates dilated convolution, channel attention, pixel attention, and takes into account the entire feature map when processing.

- Multi-head Self Channel Attention(MH-SCA) block is developed to make the processing of feature maps in the original resolution more efficient. Processing high-resolution feature maps requires more
non-local consideration, so the network is more attention-based than convolution which only considers local relationships.

The rest of this paper is organized as follows. Section II summarizes the related works in the literature. Section III presents the proposed method. Section IV compares the performance of the proposed method against the latest works in the literature, and Section V concludes the paper.

II. RELATED WORKS
A. SINGLE IMAGE DEHAZING
There are many methods for single image haze removal, which can be divided into the following two main categories. One is prior-based methods [24], [25] and the other is learning-based methods [8], [10], [26].

1) PRIOR-BASED METHODS
Prior-based methods obtain the transmission map and ambient light based on various statistical information from the image and uses them to generate a haze-free image from the scattering model Eq. (1). Dark Channel Prior(DCP) [25], creates a dark channel based on the hypothesis that the minimum value of a patch without added haze is 0 and calculates a transmission map based on the dark channel. Previous prior-based methods such as DCP can generate high quality images, but they cannot adapt to real world images, such as outdoor haze, where their assumptions are not satisfied.

2) LEARNING-BASED METHODS
With the development of deep learning, methods using deep learning have emerged in the field of haze removal. In the early studies of learning-based methods, it was difficult to obtain a clean image directly from a hazy image, so many of them estimated the ambient light and transmission map defined in the scattering model separately. DCPDN [14] succeeded in generating accurate transmission maps using edge preserving loss and achieved a very high quantitative evaluation. However, non-uniform haze images in the real world do not fulfill the following conditions in the scattering model that the attenuation rate of information due to haze depends on the distance and that the ambient light is constant in the image. For this reason, end-to-end methods that can be applied to real-world haze images are gaining more and more attention.

GridDehazeNet [9] is one of the end-to-end methods that has achieved a significant increase in quantitative evaluation compared to conventional methods. This method improves the results by combining downsampled and upsampled feature maps using attention in a grid pattern. Encoder-decoder networks such as GridDehazeNet has become mainstream, and MSBDN [12] is another one of these methods. MSBDN focuses on, the difficulty of recovering the image to its original resolution by using back projection, which has been used in super-resolution. MSBDN is able to reduce the artifacts in images. FFA-Net [10] is a different approach to produce a cleaner output. This model is deeply layered without downsampling, ensuring the expressiveness and receptive range. Moreover, by embedding attention into each layer module, the network provides more haze-aware feature extraction and greatly improves the quality. The great improvement in accuracy of FFA-Net has indicated the benefit of attention mechanisms in haze removal, and many methods have followed suit and incorporated attention mechanisms such as pixel attention and channel attention. We proposed a module with multi-head self-attention in addition to the attention used in conventional methods.

B. MULTI-HEAD SELF-ATTENTION
To explain Multi-head Self-Attention, Self Attention [27] and Multi-head Attention [27] are described in detail.

1) SELF-ATTENTION
Self-attention is based on the idea of Scaled Dot-Product Attention [27]. Scaled Dot-Product Attention is an Attention method commonly used in language models. This attention method first generates three vectors: query $Q$, key $K$ and value $V$. Given these three vectors, $Q$, $K$, and $V$, the output vector is expressed as follows.

$$ Attention(Q, K, V) = softmax(\frac{Q^T K}{\sqrt{d}}) V $$  \hspace{1cm} (2)

where $d$ denotes the number of dimensions. In Eq. (2), $Q^T K$ represents the inner product of $Q$ and $K$, and the value is calculated based on the similarity between the query and key.

Self Attention [27] is a method to get $Q$, $K$, $V$ from the same vector in Scaled Dot-Product Attention. This makes it possible to understand the relationship between words in a sentence or between pixels in an image.

2) MULTI-HEAD ATTENTION
Multi-head Attention [27] is a way of dividing the input vector into several vectors called heads, computing the Attention for each head, and finally concatenating them to obtain a new vector. The formula is as follows.

$$ Multi-head(Q, K, V) = concat(head)W^o $$ \hspace{1cm} (3)

$$ head_i = Attention(XW^Q_i, XW^K_i, XW^V_i) $$ \hspace{1cm} (4)

When $Q$, $K$, $V$ are calculated from the same vector as in Self Attention, the similarity to oneself inevitably increases, making it difficult to see the relationship between each element. Therefore, this method compresses the head to change the viewing position.

III. PROPOSED METHOD
The overall structure of the proposed method is shown in Fig. 2. The feature map generated using Mixed Encoder(ME) is processed and upsampled in a branch composed of Enhanced Feature Attention(EFA) blocks. The feature map is added to the extracted features one scale higher by...
pixel wise sum and processed in the branch of that scale. Images are generated at $i = 1, 2, 3$ through Multi Output Branches (MOB), and the output through Multi-head Self Channel Attention (MH-SCA) block at $i = 1$ is the final output image of this model.

### A. MIXED ENCODER

Mixed Encoder represents the area highlighted in red in Fig. 2. The feature extractions at the original scale and half the original scale are performed by VGG16. On the other hand, DenseNet201 [14] is used for feature extraction at 1/4 and 1/8 of the original scale. The motivation for this approach is to achieve effective feature extraction while keeping the number of parameters low.

Although many feature extractors, such as DenseNet is a very good feature extractor and is often used as an encoder in fog removal, it is not possible to extract low-level features of an image using only DenseNet because the image is downsampled by pooling and convolution with a stride of 2 at an early stage. On the other hand, VGG16 is considered superior in feature extraction at the original resolution scale. However, if all encoders are integrated with VGG16, the number of parameters becomes large due to the number of filters in the lower scales. Therefore, by using VGG16 at larger scales and then using the more powerful DenseBlock for smaller scales, we are able to perform effective feature extraction while maintaining the parameters low.

### B. ENHANCED FEATURE ATTENTION

Although introducing the attention network in the model may need more parameters, it can improve the performance. Then we introduce the attention networks by taking into account the trade-off between the number of parameters and the performance improvement.

The structure of the Enhanced Feature Attention (EFA) module is an enhancement of the Feature Attention (FA) module proposed in FFA and is shown in Fig. 3. EFA is divided into two parts: the dilation part and the attention part. The dilation part is structured using dilated convolutions to expand the receptive field of the model. To enable feature extraction over a wide range of feature sizes, dilated convolutions with dilation rates of 1, 2, 4, 8, 16 are connected in parallel. In the attention part, channel attention and pixel attention are processed in parallel, and they are concatenated and propagated to the next block by convolution. This is reasonable because both channel attention and pixel attention can see the same feature map, and it prevents one of them from becoming a bottleneck.
C. MULTI OUTPUT BRANCH
The Multi Output Branch (MOB) is highlighted in purple in Fig. 2. The MOB employs the refinement module [8] for outputs in each scale. In the Refinement module (Fig. 4), the input feature maps are first downsampled by a factor of 4, 8, 16, 32, and then each feature map is compressed in the channel dimension. Then, each feature map is upsampled to the original size, and these compressed channels and the input feature map are concatenated. Finally, it is passed through a convolution layer to output a dehazed image. This module allows us to obtain an output in which more global information is taken into account. The shallow layer features are combined with the deep layer features by a skip connection. The concatenated features are propagated to higher-scale branches.

\[ \text{output} = \text{conv}(\text{upsample}(\text{concatenate}(\text{input}, \text{compressed channels}))) \]

\[ \text{concatenate}(\text{input}, \text{compressed channels}) \]

D. MULTI-HEAD SELF CHANNEL ATTENTION BLOCK
Multi-head Self Channel Attention (MH-SCA) block is the yellow block that is embedded in the \( i = 1 \) scale branch in Fig. 2. In the original scale part of the branch, it is difficult to capture the entire feature map using only convolution layers due to the high resolution. Therefore, instead of using a convolution, an attention-based structure is used to view more global information. We adopt the MH-SCA block as the attention block. This block is composed of MH-SCA and Feed-Forward Network (FFN) [27] as shown in Fig. 5. The FFN contains a layer normalization, a convolution layer with a kernel size of 1, and a GELU [28], and has the effect of reshaping the channel directional positioning that has been broken by the Multi-head Attention.

\[ \tilde{z} = \text{softmax}(q^T k v) \]  
\[ \tilde{z}_c = F_{\text{scale}}(x_c, \tilde{z}_c) = \tilde{z}_c x_c \]

\[ X = [x_1, x_2, \ldots, x_c], \text{a new feature map } \tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_c] \]

1) MULTI-HEAD SELF CHANNEL ATTENTION
Self Channel Attention (SCA) introduces the concept of Self Attention to conventional Channel Attention. In conventional channel attention, a feature map is processed by Global Average Pooling (GAP), and the resulting vector is passed through fully connected layers to generate a new vector as weights for each channel. If all channels are considered to determine the attention weights, then it can negatively affect the unrelated channels because some channels have correlations with each other while others do not. By using Self Attention on the vectorized feature map by Global Average Pooling as shown in Fig. 6, it is possible to consider the relationship between channels. The feature map is updated according to the following formula.

- First, Global Average Pooling is performed on the feature map \( X \), and query \( q \), key \( k \), and value \( v \) are generated from the vectors.

\[ z = \text{GAP}(X) \]
\[ q = w_q z \]
\[ k = w_k z \]
\[ v = w_v z \]

(5)

Here, GAP stands for Global Average Pooling, and when \( X \) is a feature map of \( H \times W \times C \), \( z \) is a one-dimensional vector of \( 1 \times 1 \times C \).

- A new vector \( \tilde{z} \) is updated from the obtained \( q \), \( k \), and \( v \) as follows according to the concept of Self Attention.

\[ \tilde{z} = \text{softmax}(q^T k v) \]  
\[ \tilde{z}_c = F_{\text{scale}}(x_c, \tilde{z}_c) = \tilde{z}_c x_c \]  
\[ X = [x_1, x_2, \ldots, x_c], \text{a new feature map } \tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_c] \]

\[ \tilde{X} = \text{concatenate}(\text{input}, \tilde{X}) \]

\[ \text{concatenate}(\text{input}, \tilde{X}) \]

- From the new vector \( \tilde{z} \) and feature map \( X = [x_1, x_2, \ldots, x_c] \), a new feature map \( \tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_c] \) is generated according to the following equation

\[ \tilde{x}_c = F_{\text{scale}}(x_c, \tilde{z}_c) = \tilde{z}_c x_c \]  

\[ F_{\text{scale}} \] is the element wise product of the scalar \( \tilde{z}_c \) and the feature map \( x_c \in \mathbb{R}^{H \times W} \), where \( \tilde{z}_c \) is \( c \)-th element of \( \tilde{z} \).
Furthermore, by making SCA a multi-head structure, a more effective self-attention is achieved.

E. LOSS FUNCTION

The overall loss function is expressed as

$$L = \sum_{i=1}^{3} \lambda_i(L_p^i + L_{ms}^i + L_s^i)$$  

(8)

where \(L_p\), \(L_s\), and \(L_{ms}\) represent the perceptual loss, smooth \(L_1\) loss, and Multi-scale Structural Similarity loss respectively. \(i\) denotes each scale, and losses are weighted by \(\lambda_i\) according to the scale and summed. With this loss, the MOB structure can improve the accuracy from low-scale to high-scale branches to reduce the deterioration.

1) PERCEPTUAL LOSS

Perceptual loss takes the loss for each multi-scale feature map extracted by a pre-trained network. In this method, VGG16 is used for the pre-trained network, and the perceptual loss is expressed as follows.

$$L_p^i = \frac{1}{C_J H_J W_J} \| \phi_j(\hat{J}) - \phi_j(J) \|^2_2$$  

(9)

where \(\phi_j(\hat{J})\) represents the VGG16 feature maps from the output image and \(C_J\), \(H_J\) and \(W_J\) indicates the number of channels, height and width of each feature map respectively.

2) SMOOTH \(L_1\) LOSS

Smooth \(L_1\) loss \(L_s\) uses the \(L_1\) norm, which is less sensitive to outliers than the MSE loss. In addition, the gradient is made smoother, making it less prone to gradient explosion, and is expressed as follows [29].

$$L_s^i = F_s(\hat{J}(x) - J(x))$$  

(10)

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

(11)

3) MULTI-SCALE STRUCTURAL SIMILARITY LOSS

Since the optimal scale for evaluation varies depending on the image, Multi-scale Structural Similarity [30][MS-SSIM] was proposed to take into account SSIM at various scales. First, to obtain multi-scale information, the image pair is low-pass filtered and downsampled \(M\) times by a factor of 2. Next, the contrast \(c\) and structure \(s\) are compared, and finally the luminance \(l\) is measured.

All components are multiplied together to compute MS-SSIM, which is expressed in the following equation.

$$\text{SSIM}(x, y) = \left[\text{M}(x, y)\right]^{\alpha} \cdot \prod_{j=1}^{M} \left[C_j(x, y)\right]^{\beta_j} \cdot \left[S_j(x, y)\right]^{\gamma_j}$$  

(12)

The MS-SSIM loss is the loss between the MS-SSIM of the output and the ground truth, and has a maximum value of 1. MS-SSIM loss is expressed in the following equation.

$$L_{ms} = 1 - \text{SSIM}(x, y)$$  

(13)

where \(\alpha, \beta_j, \gamma_j\) are the default parameters computed by subject experiments.

IV. EXPERIMENTAL RESULTS

In this section, we conduct an ablation study to demonstrate the effectiveness of each of the proposed components by constructing models with and without these modules. We also compared our method quantitatively and qualitatively with conventional dehazing methods.

A. EXPERIMENT SETTINGS

1) DATASETS

In this study, we used NH-HAZE dataset [18] for the ablation study, and RESIDE [31], Dense-HAZE [19], and NH-HAZE datasets for the comparison with other conventional methods. RESIDE, proposed in 2018 by Li et al. is mainly used as a synthetic image dataset. The large number of images allows for training on large datasets in deep learning-based dehazing tasks. It is divided into two main parts, the training dataset and the testing dataset. In this study, we use the Indoor Training Set(ITS) and Synthetic Objective Testing Set(SOTS) indoor datasets, which are most often used in single image haze removal, as training and testing dataset, respectively. Dense-HAZE and NH-HAZE are both real-world datasets consisting of hazy, clean pairs. The Dense-HAZE dataset has uniform dense haze over its images, while NH-HAZE is a dataset with inhomogeneous haze. Both of these datasets consist of 40 training data, 5 validation data, and 5 test data. In our experiments, we used 40 training data and 5 validation data.

2) IMPLEMENTATION

EMSAN is implemented by Pytorch with a RTX 1080Ti GPU. Images are randomly rotated 90°, 180°, and 270°, horizontal flipped, and patches of 256 × 256 are randomly cropped to generate training samples. We adopt the Adam optimizer with a batch size of 4, and the learning rate is adjusted using the cosine learning rate decay [10], [32]. Other hyper parameters are set as \(\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}\). Adjusting parameters in Eq.(8) are set as \(\lambda_1 = 2, \lambda_2 = 1\) and \(\lambda_3 = 4\).

B. ABLATION STUDY

In order to investigate the effectiveness of the various architectures proposed in this study, we carried out an ablation study using the NH-HAZE dataset. In addition to the base model(Fig. 7a) and the proposed network(Fig. 2), three other

| Parameters | Base Model | Model-1 | Model-2 | Model-3 | Ours |
|------------|------------|---------|---------|---------|------|
| Beta alpha | 10.77M     | 10.98M  | 11.47M  | 5.40M   | 4.76M|
| PSNR       | 20.16      | 20.62   | 20.92   | 21.26   | 21.40|
| SSIM       | 0.7188     | 0.7256  | 0.7262  | 0.7268  | 0.7333|

TABLE 1. Ablation study.
models were developed and tested, as shown in Fig. 7, to better demonstrate the effects of each component.

The base model is shown in Fig. 7a. The encoder is a pre-trained VGG16, and the decoder is block and group structures with the FA module developed in FFA-Net [10]. Model-1 (Fig. 7b) is a model in which the FA module of the base model is changed to the Enhanced FA module. Model-2 (Fig. 7c) adds the Multi-Output Branch. Model-3 (Fig. 7d) replaces the encoder of model-2 with a Mixed Encoder. Finally, the block of the original resolution branch in model-3 is changed to MH-SCA block.

The results of experiments using these models are shown in Table 1. As Table 1 shows, the addition of each component improves the performance. The Mixed Encoder and MH-SCA block also reduce the number of parameters, indicating that our method enables effective processing.

C. HAZE REMOVAL EVALUATIONS
Table 2 shows the comparison with other dehazing methods on RESIDE [31], NH-HAZE [18], Dense-HAZE [19]. In all of the datasets, our method outperforms the other methods in PSNR. Although the number of parameters do not reach the level of AECR-Net [33], our method is superior to AECR-Net in PSNR and SSIM. This prove to be effective against the other methods in terms of correlation with numerical evaluation.

We also show visual comparisons with SOTA methods in Fig. 8, Fig. 9 and Fig. 10. Although there is not much difference in appearance in the synthetic image dataset SOTS as indicated by Fig. 8, the proposed method is able to recover objects in distant areas with dense haze and similar in color to the haze.

Fig. 9 shows the results of the real image dataset Dense-HAZE. As can be seen from Fig. 9, most of the methods are not able to produces high quality images due to the dense haze, but EMSAN is the closest to the ground truth in terms of the object structure. This shows that ME successfully extracts the background features.

More accurate haze feature extraction is important for the real image dataset NH-HAZE because of the non-uniformity of haze distribution. GridDehazeNet [9] has a narrow receptive field and can only consider local information since it is composed of simple convolution layers and only down-samples twice. This results in the mottled patterns in the output images in Fig. 10. On the other hand, MSBDN [12] downsamples deeply to obtain global information of the entire image, so the mottled patterns caused by the locality of the network disappears. However, the image degradation is...
FIGURE 8. Visual comparison on SOTS datasets.
obvious because of the excessive reduction in spatial resolution. FFA-Net [10] does not downsample, so the quality of fine edges such as textured leaves is less reduced than in MSBDN, but the deeper layers still do not solve the problem of CNN localization, causing mottled artifacts. Although AECR-Net is the most effective in removing haze among the previous methods, inaccuracies in understanding haze information with few parameters are noticeable, such as removing too much haze or, conversely, not enough.

It can be seen from our output images that these problems observed in conventional methods have been solved by EMSAN.

V. CONCLUSION

In this paper, we propose a single image haze removal model that uses Attention to process multi-scale features. In order to cope with the problems of conventional methods, such as image degradation, the increase in the number of parameters, and the difficulty in expanding the receptive field of the network, we have developed a Mixed Encoder, an Enhanced Feature Attention module, a Multi Output Branch, and Multi-Head Self Channel Attention Block. By devising this new network structure, the proposed EMSAN is able to achieve higher quantitative and qualitative evaluations than previous methods. Future work is to explore more lightweight dehazing networks while improving performance.

REFERENCES

[1] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, “Residual conv-deconv grid network for semantic segmentation,” 2017, arXiv:1707.07958.
[2] A. Kirillov, R. Girshick, K. He, and P. Dollar, “Panoptic feature pyramid networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6399–6408.
[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3339–3348.
[4] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, “End-to-end united video dehazing and detection,” in Proc. AAAI, Apr. 2018, pp. 1–8.
[5] S. G. Narasimhan and S. K. Nayar, “Chromatic framework for vision in bad weather,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2000, pp. 598–605.
[6] R. Hide, “Optics of the atmosphere: Scattering by molecules and particles,” Phys. Bull., vol. 28, no. 11, p. 521, Nov. 1977.
[7] S. G. Narasimhan and S. K. Nayar, “Vision and the atmosphere,” Int. J. Comput. Vis., vol. 48, no. 3, pp. 233–254, 2002.
[8] Y. Qu, Y. Chen, J. Huang, and Y. Xie, “Enhanced pix2pix dehazing network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 8160–8168.
[9] X. Liu, Y. Ma, Z. Shi, and J. Chen, “GridDehazeNet: Attention-based multi-scale network for image dehazing,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7314–7323.
[10] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, “FFA-Net: Feature fusion attention network for single image dehazing,” in Proc. AAAI Conf. Artif. Intell., Apr. 2020, pp. 11908–11915.
[11] M. Hong, Y. Xie, C. Li, and Y. Qu, “Distilling image dehazing with heterogeneous task imitation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3462–3471.

[12] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, “Multi-scale boosted dehazing network with dense feature fusion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2157–2167.

[13] L. Yan, W. Zheng, C. Gou, and F.-Y. Wang, “Feature aggregation attention network for single image dehazing,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2020, pp. 923–927.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[15] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “DehazeNet: An end-to-end system for single image haze removal,” IEEE Trans. Image Process., vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[16] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, “Domain adaptation for image dehazing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2808–2817.

[17] L.-A. Tran, S. Moon, and D.-C. Park, “A novel encoder-decoder network with guided transmission map for single image dehazing,” 2022, arXiv:2202.04757.

[18] C. O. Ancuti, C. Ancuti, and R. Timofte, “NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2020, pp. 444–445.

[19] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, “Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 1014–1018.

[20] R. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent., 2015, pp. 234–241.

[21] S. Dai, M. Han, Y. Wu, and Y. Gong, “Bilateral back-projection for single image super resolution,” in Proc. IEEE Multimedia Exp. Int. Conf., Jul. 2007, pp. 1039–1042.

[22] M. Irani and S. Peleg, “Improving resolution by image registration,” CVGIP: Graph. Models Image Process., vol. 53, no. 3, pp. 231–239, May 1991.

[23] M. Haris, G. Shakharovitch, and N. Ukita, “Deep back-projection networks for super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1664–1673.

[24] X. Fan, Y. Wang, X. Tang, R. Gao, and Z. Luo, “Two-layer Gaussian process regression with example selection for image dehazing,” IEEE Trans. Circuits Syst. Video Technolog., vol. 27, no. 12, pp. 2505–2517, Dec. 2017.

[25] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2341–2353, Sep. 2010.

[26] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, “AOD-Net: All-in-one dehazing network,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4770–4778.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[28] D. Hendrycks and K. Gimpel, “Bridging nonlinearities and stochastic regularizers with Gaussian error linear units,” 2016, arXiv:1606.08415.

[29] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[30] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, Jul. 2003, pp. 1398–1402.

[31] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, “Benchmarking single image dehazing and beyond,” 2017, arXiv:1712.04143.

[32] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of tricks for image classification with convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 558–567.

[33] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, “Contrastive learning for compact single image dehazing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10551–10560.

[34] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

WATARU IMAI (Member, IEEE) received the B.E. degree in electrical engineering from Keio University, Yokohama, Japan, in 2020, where he is currently pursuing the M.E. degree, under the supervision of Prof. Masaaki Ikehara. His research interests include the area of image restoration, image dehazing, and neural networks.

MASAAKI IKEHARA (Senior Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1984, 1986, and 1989, respectively. He was appointed as a Lecturer at Nagasaki University, Nagasaki, Japan, from 1989 to 1992. In 1992, he joined the Faculty of Engineering, Keio University. From 1996 to 1998, he was a Visiting Researcher with the University of Wisconsin–Madison, Madison, and Boston University, Boston, MA, USA. He is currently a Full Professor with the Department of Electronics and Electrical Engineering, Keio University. His research interests include the areas of multi-rate signal processing, wavelet image coding, and filter design problems.

* * *