Trident Dehazing Network

Jing Liu[1] Haiyan Wu[1] Yuan Xie[1]*Yanyun Qu[2] Lizhuang Ma[1]
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]School of Information Science and Engineering, Xiamen University, Fujian, China
51174500035@stu.ecnu.edu.cn, 704289013@qq.com,
xieyuan8589@foxmail.com, yyqu@xmu.edu.cn, lzma@cs.ecnu.edu.cn

Abstract

Most existing dehazing methods are not robust to nonhomogeneous haze. Meanwhile, the information of dense haze region is usually unknown and hard to estimate, leading to blurry in dehaze result for those regions. Focusing on these two issues, we propose a novel coarse-to-fine model, namely Trident Dehazing Network (TDN), to learn the hazy to hazy-free image mapping with automatic haze density recognition. In detail, TDN is composed of three sub-nets: the Encoder-Decoder Net (EDN) is the main net of TDN to reconstruct the coarse hazy-free feature; the Detail Refinement sub-Net (DRN) helps to refine the high frequency details that was easily lost in the pooling layers in the encoder; and the Haze Density Map Generation sub-Net (HDMGN) can automatically distinguish the thick haze region with thin one, to prevent over-dehazing or under-dehazing in regions of different haze density. Moreover, we propose a frequency domain loss function to make supervision of different frequency band more uniform. Extensive experimental results on synthetic and real datasets demonstrate that our proposed TDN outperforms the state-of-the-arts with better fidelity and perceptual, generalizing well on both dense haze and nonhomogeneous haze scene. ***Our method won the first place in NTIRE2020 nonhomogeneous dehazing challenge.***

1. Introduction

Haze is one primary source of image degradation, which results in the low contrast, color distortion and blurring problems. Previous works [21, 24] have established that haze heavily affects the performance of the high-level tasks, such as classification or semantic segmentation, and thus a preprocessing that recovers back the depicted subjects is highly desirable. To remove haze and improve visibility of the hazy image, many dehazing methods [17, 37, 29, 10, 26, 11, 20, 14, 21, 23, 27, 35] have been proposed.

In early atmosphere scattering model based dehazing methods such as dark-channel prior [17], contrast color-lines [15] and haze-line prior [10], the global atmospheric light and the medium transmission map are evaluated by hand-crafted priors. These priors are supposed to distinguish between hazy region and hazy free region. Some learning-based methods [29, 26, 35, 21, 11] also follow the conventional procedure of dehazing: estimate the transmission map and the atmospheric light, and recover the hazy-free image based on the atmospheric scattering model. However, there are many non-uniform weather conditions, where neither the haze nor the transmission map can be accurately estimated, which in turn affects the subsequent dehazing, resulting in over-colored and lack-of-details undesirable results. Existing methods face challenges in dense haze and nonhomogeneous haze conditions. In order to get over the parameter limitation brought by the physical model, an end to end model that directly learn the input hazy images to the clear hazy free images mapping without drawing support from the physical model is needed.

Because it is difficult to obtain the image pairs of haze image and corresponding ground truth at the same time, the number of dehazing training sets in real scenes is very limited, most works use synthetic datasets as training sets. However, the difference between the synthetic datasets generated by the synthetic algorithm and the real haze scene is large, and the sample size of the real haze scene training set is too small, so we need more accurate and stronger prior to integrate the additional information into the deep dehazing model. Encoder-Decoder architecture is a good choice. The backbone pretrained on the massive clear image set like Imagenet can be used as the encoder, more accurate and real prior information is integrated into the dehazing model, and therefore our proposed Trident Dehazing Network (TDN) introduces Encoder-Decoder architecture as the main net to reconstruct the coarse hazy free image. Moreover, a pre-trained encoder can also greatly speed up the convergence of the network.

However, due to the unlearnable pooling layers and

*Corresponding author

downsampling layers in most backbones (such as ResNet, DenseNet, and etc.), high frequency details are lost, and region of reconstructed image where there is dense haze in the input hazy image is very blurry. To solve this problem and get reconstructed hazy free images with better perceptual quality, we introduce a Details Refinement sub-Net (DRN) to help the main net getting sharper and more faithful result. We also produce a novel FFT loss function to supervise the frequency domain information. With the FFT loss, supervise for different frequency bands are more uniform, and high frequency information is easier captured. Moreover, to help adapting nonhomogeneous haze scene and prevent over-dehazing or under-dehazing in regions of different haze density, we add a U-Net style Haze Density Map Generation sub-Net (HDMGN) to learn the input hazy image to haze density map mapping automatically. With these two sub-nets, our proposed Trident Dehazing Network can not only adapt the dense haze scene, but also adapt the nonhomogeneous haze scene. Extensive ablation experiments show the effectiveness of the three sub-nets in proposed TDN and FFT loss. Experiments comparing the PSNR, SSIM and LPIPS metrics with previous state-of-the-art methods on widely-used dehazing benchmark RESIDE and real world NTIRE dehazing challenges test sets demonstrate that TDN surpasses all the previous methods with better fidelity and perceptual.

Overall, our contributions are three-folds as below:

- We propose a novel FFT loss function to supervise the frequency domain signal. Combining spatial domain loss and frequency domain loss, supervise for different frequency bands are more uniform, and high frequency information is easier captured.

- We propose a novel coarse to fine model Trident Dehazing Network (TDN) to end to end learn the hazy to hazy free image mapping with automatic haze density recognition. Our proposed Details Refinement sub-Net can refine the coarse feature maps with more high frequency details, and our proposed Haze Density Map Generation sub-Net can automatically reconstruct the haze density map with no extra supervision for nonhomogeneous dehazing.

- **Our proposed TDN won the first place in NTIRE2020 nonhomogeneous dehazing challenge.** Extensive experiment results demonstrate on commonly used dehazing benchmarks that TDN surpasses the previous state-of-the-art methods with better fidelity and perceptual on not only synthetic datasets but also real-world hazy scene, and it's also able to generalize well on both dense haze and nonhomogeneous haze scene.

2. Trident Dehazing Network

In this section, we first overview the proposed Trident Dehazing Network (TDN), then we introduce some novel loss functions used for training TDN, which may be useful for other dehazing networks.

2.1. Network Architecture

In this paper, we proposed a Trident Dehazing Network (TDN) to directly learn a mapping from the input real world nonhomogeneous hazy image to the hazy-free clear image. As shown in Figure 1, TDN consists of three sub-nets, the Encoder-Decoder sub-Net (EDN), the Details Refinement sub-Net (DRN), and the Haze Density Map Generation sub-Net (HDMGN), each of which is used for a specific purpose: EDN reconstructs the coarse features of hazy-free images, DRN complements the high frequency details of the hazy free image features, and HDMGN helps obtaining the density of haze in the different region of the input hazy image. The deformable [38] convolution block gets the final clear output from the concatenated feature maps of three sub-nets.

Encoder-Decoder sub-Net. In the proposed EDN, the encoder part supports ResNet-Style backbones, such as ResNet [18], ResNext [32], Res2Net [16], DPN [13], and etc.. It should contain the head layers, which support $4\times$ downsampling and shallow feature extracting, one layer without downsampling, and three layers with downsampling operators in the first bottleneck. We use DPN92 pretrained in ImageNet1K as the backbone of our encoder part. It's a powerful feature extractor, where the extracted shallow features represent low level visual features and the extracted deep features are capable to capture the semantic information. EDN fixes the encoder during training to exploit the power of pretrained DPN as "priors", and the encoder brings a lot of extra information for small sample training set (such as Dense Haze, NH-Haze, and etc.).

The decoder is composed of five Deformable Upsampling Blocks (DUB), as shown in Figure 1 (bottom right). The input feature is first fed into a 3×3 deformable convolution block, and then concatenated with the output features. The concatenated features are fed into an 1×1 deformable convolution block and an nearest-upsampling $2\times$ layer to get the upsampled features as the input features of the next DUB.

We add skip connections from the output of the first downsampling block in layer 2 and that in layer 3 to the input of DUB 2, 3 by concatenating (cat) the feature maps, respectively. We use trainable instance normalization [30] for skip connections. Our Encoder-Decoder sub-Net has large capacity, and skip connections make the information smoothly flow to easily train a large network.

Haze Density Map Generation sub-Net. As shown in Figure 2, we use a simple U-Net architecture proposed in pix2pix [19] network to achieve haze density map generation. Different with U-Net in pix2pix network, we add a tail 3×3

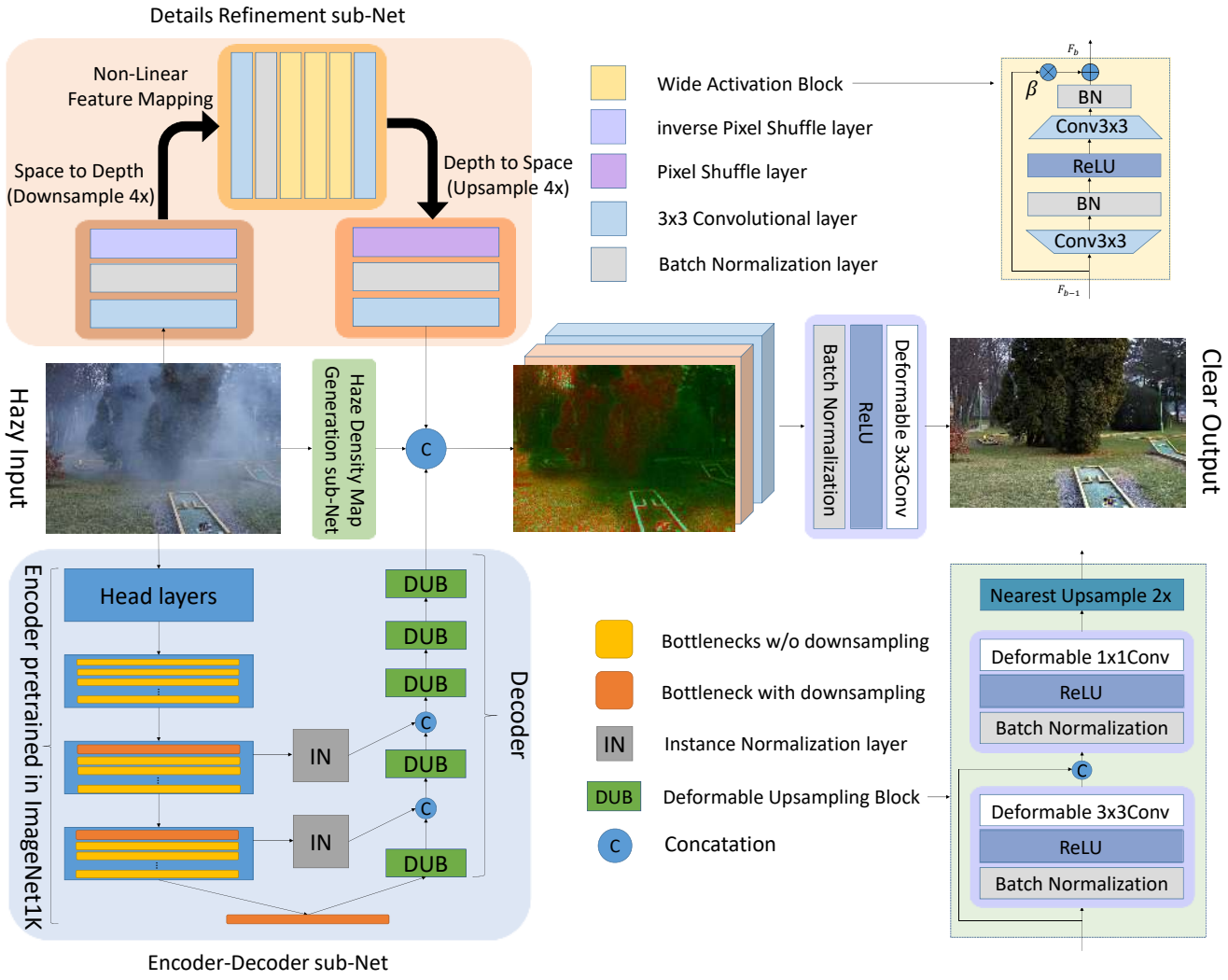

Figure 1: The architecture of the proposed Trident Dehazing Network (TDN), the Details Refinement sub-Net (DRN) and the Encoder-Decoder sub-Net (EDN). $\oplus$ represents tensor addition and $\otimes$ represents tensor multiplication respectively. TDN consists of three sub-nets: EDN, DRN and HDMGN. The haze density maps and intermediate feature maps output by three sub-nets are then concatenated and fed into the tail deformable [38] convolution block to get the clear output.

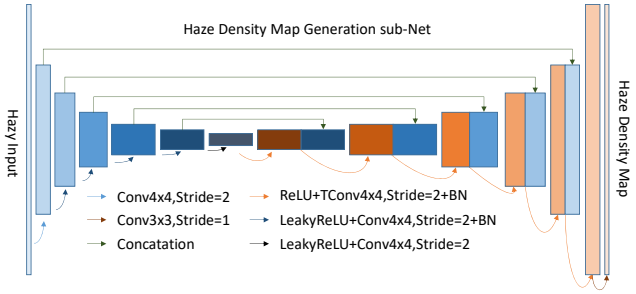

Figure 2: The architecture of the proposed Haze Density Map Generation sub-Net. "TConv" is the abbreviation of "Transpose Convolution".

convolutional layer to refine the output. Due to the size division requirement, there are only 6 downsampling and upsampling operators in the U-Net, and the input size should be divisible by 64. As shown in Figure 3, the greener the region in the visualization haze density map is, the more haze there is. Note that we don't use any extra supervision to the haze density map. The U-Net can automatically achieve it.

Details Refinement sub-Net. Inspired by pre-upsampling based single image super resolution networks, we propose a Details Refinement sub-Net to do the non-linear feature mapping on downsampled $4\times$ factor. Inverse Pixel Shuffle layer is used to change the feature maps from spatial to depth (downsampling/desubpixel), and Pixel Shuffle layer [28] is used to change the feature maps from depth to spatial (upsampling/subpixel). As shown in Figure 1, three Wide Activation Blocks (WAB) provide the non-linear fea-

ture mapping on 4× downsampled factor. In the WAB, there are two 3×3 convolutional layers (followed by batch normalization layer) and a wide activation layer proposed in [33]. The channel expand factor of WAB is 4. Motivated by [31], we use residual scaling, *i.e.*, scaling down the residuals by multiplying a constant between 0 and 1 before adding them to the main path, preventing training-instability. Adding the Details Refinement sub-Net, the training process is more stable and the final output can be enhanced from a somewhat blurry reconstruction result to a sharper one with more clear details.

2.2. Loss functions

Given the hazy image and the ground-truth haze-free image, we intend to learn network parameters by minimizing a loss function that consists of three different components, each of which is used for a specific purpose. The individual loss terms are described as follows:

Spatial Domain Loss: This is a standard ℓ_1 loss function commonly used in image reconstruction tasks. It supervises the reconstruction hazy free results in the spatial domain:

$$L_s = \frac{1}{N}\sum_{i=1}^{n} \|x_i^{gt} - TDN(x_i^{hazy})\|_1, \tag{1}$$

where x_i^{gt}, $TDN(x_i^{hazy})$ denote the i-th ground truth sample and hazy free sample reconstructed by the proposed TDN, respectively.

Frequency Domain Loss: This is a novel loss function which supervises the reconstruction hazy free results in the frequency domain. The output is the amplitude and phase in the frequency domain, and Fast Fourier transform (FFT) loss supervise both the amplitude and phase by ℓ_1 loss function:

$$A_{x_i^{gt}}, P_{x_i^{gt}} = FFT(x_i^{gt}), \tag{2}$$

$$A_{TDN(x_i^{hazy})}, P_{TDN(x_i^{hazy})} = FFT(TDN(x_i^{hazy})), \tag{3}$$

$$\begin{aligned} L_f = \frac{1}{N}\sum_{i=1}^{n}(&\|A_{x_i^{gt}} - A_{TDN(x_i^{hazy})}\|_1 \\ &+\alpha\|P_{x_i^{gt}} - P_{TDN(x_i^{hazy})}\|_1), \end{aligned} \tag{4}$$

where A_i and P_i refer to the amplitude and phase of the i-th image sample respectively, FFT($\cdot$) denotes the fast fourier transform and α serves as the trade-off parameter between the two terms (in our experiment, $\alpha = 1$). For the implementation, we use PyTorch where the FFT operator can be implemented by torch.fft and easily calculated in GPU. Note that FFT brings a very limited time cost in the training process, while the proposed frequency domain loss greatly improves the visual perceptual quality of the reconstructed image without any inference cost.

Threshold Limitation Loss: The input and output range of TDN is from 0 to 1. The output directly through TDN may result in threshold overflow. We use BReLU layer [11] in the tail of TDN to limit the output data into range [0,1]:

$$O_i = B(TDN(x_i^{hazy})), \tag{5}$$

where $B(\cdot)$ refers to the BReLU operator and O_i refers to the final output of the i-th image sample, respectively. The proposed threshold limitation loss can be formulated as:

$$L_l = \frac{1}{N}\sum_{i=1}^{n} log(\|O_i - TDN(x_i^{hazy})\|_1 + 1). \tag{6}$$

The region of black and white pixels in ground truth is more easily reconstructed by dehazing networks to overflow the threshold, they can be fixed by the BReLU layer. But there are other value which will overflow. By minimizing the threshold limitation loss, these outliers can be corrected to the normal range.

The **overall loss function** is defined as:

$$L = \alpha L_s + \beta L_f + \gamma L_l, \tag{7}$$

where α, β and γ serve as the trade-off parameters to balance different loss terms.

3. Experiments

In this section, we start from describing the experimental settings, *i.e.,* datasets, evaluation metrics and implementation details, and then give the ablation study and model analysis respectively. Finally, we perform the experiments to evaluate the performance of ours as well as competitors.

3.1. Experimental Settings

Datasets. In this paper, we evaluate our network on two types of datasets: synthetic dataset and real-world dataset. For synthetic data, we choose a well-known and representative RESIDE [22] dataset. For real-world data, we experiment with the I-HAZE dataset [7], O-HAZE dataset [6] which are used in NTIRE2018 Dehazing Challenge [1], Dense Haze dataset [3] used in NTIRE2019 Dehazing Challenge [2], and NH-Haze dataset [8, 4] used in NTIRE2020 Dehazing Challenge [5, 9].

As a standard benchmark widely used in dehazing task, RESIDE consists of both indoor and outdoor, both synthetic and real-world hazy images. There are totally five subsets in it: Indoor Training Set (ITS), Synthetic Objective Testing Set (SOTS), Hybrid Subjective Testing Set (HSTS), Outdoor Training Set (OTS) and Real world Task-driven Testing Set (RTTS). The atmospheric scattering model is used, where atmospheric lights is randomly chosen between $(0.7, 1.0)$ for each channel, and scattering coefficient is randomly selected between $(0.6, 1.8)$. In our work, we use ITS which consists of 13990 synthetic images to train our network. The indoor part of SOTS is employed as our testing set, which

includes 500 indoor images. In addition, we implement our method on real-world dehazing benchmarks introduced in NTIRE2018, NTIRE2019 and NTIRE2020 dehazing challenges for further evaluation. I-HAZE, O-HAZE, Dense-Haze and NH-Haze contain 25 indoor hazy images, 35 outdoor hazy images, 45 dense hazy images and 45 nonhomogeneous hazy images and their corresponding ground truth (gt) for training respectively, 5 hazy-gt pairs for validation and 5 for testing. The NTIRE challenges datasets are captured in presence or absence of haze in various indoor and outdoor scenes using a professional haze/fog generator that imitates the real conditions of haze scenes. The details are shown in Table 1. For I-HAZE, O-HAZE and Dense-Haze datasets, we use the training set for training and validation set for testing. For NH-Haze dataset, because the validation set is not public now, we use image $1 \sim 40$ as training set and $41 \sim 45$ as testing set for ablation studies.

datasets	quantity	scenes	r	image size	format
RESIDE (ITS)	13990+500	Indoor	×	620×460	PNG
I-HAZE	25+5+5*	Indoor	✓	4500×2800	JPEG
O-HAZE	35+5+5*	Outdoor	✓	2500×2500	JPEG
Dense-Haze	45+5+5*	22I+33O	✓	1600×1200	PNG
NH-Haze	45+5*+5*	Outdoor	✓	1600×1200	PNG

Table 1: The details of dehazing benchmarks used in the paper. "I" and "O" refer to "Indoor" and "Outdoor" respectively. "*" denotes that the ground truth images are not public. The image size of I-HAZE and O-HAZE is an average estimated value. "r" denotes whether the hazy images are real world images or synthetic images.

Evaluation Metrics. For qualitative evaluation, we measure the result of our method in terms of two evaluation metrics: the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity index (SSIM), which are often used as criteria for evaluating image quality in low-level vision tasks. In addition, because the aim of NTIRE2020 dehazing challenge is to produce high quality results with the best perceptual quality and similar to the reference ground truth, we add the perceptual measures Learned Perceptual Image Patch Similarity (LPIPS) [36] for NH-Haze dataset and compare our dehazing effects with the subjective visual effects.

Model Settings and Implementation Details. During the training, patches of size 256 × 256 are cropped from the training images. We train TDN in RGB channels and augment the training dataset with randomly rotated by 90, 180, 270 degrees and horizontal flip. Our models are optimized using the Adam optimizer with β_1 = 0.9, β_2 = 0.999, and ϵ = 10^{-8} to minimize the loss functions. The initial learning rate is set to 10^{-4} and then decreases to half in 30^{th}, 55^{th} and 80^{th} epoch. The batch size is 40, the total epoch number is 100, and there are 400 iterations for each epoch. We use PyTorch to implement our models with two GTX1080Ti for training and one for testing. The training time is about one day, and the inference time of one 1600×1200 image in one GTX1080Ti is 0.64s. The backbone we use in TDN, pretrained DPN92, is first pretrained in ImageNet5K, and then fine tuned in ImageNet1K. Other parameters of TDN are initialized with random weights. The learning rate of head layers, layer1 and layer2 of DPN92 are half that of the whole network, the learning rate of layer3, layer4-block1 of DPN92 and deformable convolution offset mask layers are 0.7, 0.9 and 0.3 times that of the whole network respectively. The residual scale used in wide activation block is 0.2. We empirically set the loss weighting factors α, β and γ as 0.5, 0.5 and 1.

3.2. Ablation Study and Model Analysis

Effectiveness of three sub-nets. Due to the pooling layers in the encoder part of Encoder-Decoder sub-Net (EDN), EDN is more inclined to learn the coarse low frequency information getting somewhat blurry result. Detail Refinement sub-Net (DRN) is proposed to refine the high frequency details by the non-linear feature mapping in downsampled 4× factor. In order to adapt nonhomogeneous images, we add a U-Net style architecture to learn the hazy input image to haze density map mapping. In short, Trident Dehazing Network is a coarse to fine architecture with automatic haze density recognition for nonhomogeneous dehazing.

We provide the visualization maps of the output of three sub-nets to demonstrate the effectiveness of three sub-nets. As shown in Figure 3, the output feature map of EDN reconstructs the coarse features of the hazy free image, which contains the general outline information of the image. Note that some extreme outlier pixels cause there are some artifacts in the EDN visualization map. It will be fixed by the tail deformable convolutional layer. The tail deformable convolutional layer can automatically ignore these outliers and choose the suitable pixels by the learnable offset to do the convolution operator. HDMGN reconstructs the distribution of haze density accurately. In the output feature map of DRN, the edge information and high frequency details are the main components. The ablation study results of three sub-nets are shown in Table 2 (5)-(8). A more quickly inferenced backbone DPN68 is used in ablation studies. Setting (5) is the final setting of EDN and the whole network. Only the main-net EDN is in model (5). Model (6) and model (7) lose HDMGN and DRN respectively, and model (8) contain all sub-nets. If any sub-net is removed, the performance will be worse.

Effectiveness of Deformable Convolutional Layers. As shown in Figure 4 left, without deformable convolution, local region is recognized as the same level of haze density, which results in that when some objects become the foreground area while other parts blocked by haze become the background region, the haze part cannot be reconstructed, while the foreground part is over sharpened.

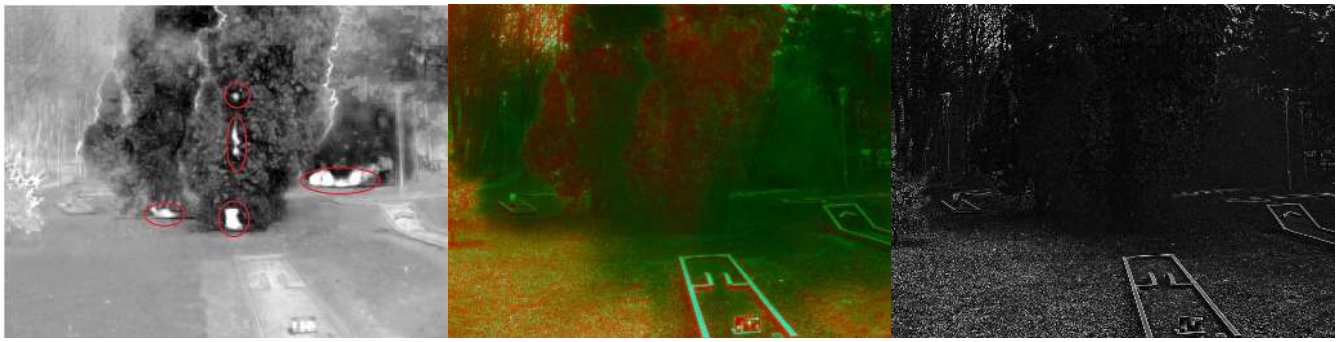

Figure 3: From left to right are the visualization maps of Encoder-Decoder sub-Net (EDN), Haze Density Map Generation sub-Net (HDMGN) and Detail Refinement sub-Net (DRN). The output of HDMGN is 3 channel feature maps, which is visualized as RGB maps after [0,1] clamped. To visualize the output of EDN and DRN, we sum the output feature maps of EDN and DRN in the channel dimension and use a sigmoid layer to get the visualization maps. The input hazy image "52.png" can be found in Figure 1.

loss function	architecture	PSNR	SSIM	LPIPS
(1)L1	DPN68	19.70	0.6260	0.507
(2)L1+BReLU	DPN68	19.87	0.6320	0.504
(3)L1+FFT+BReLU	DPN68	20.08	0.6651	0.432
(4)L1+FFT+BReLU	DPN68+IN	20.09	0.6728	0.425
(5)L1+FFT+BReLU	DPN68+IN+DC	20.19	0.6852	0.393
(6)L1+FFT+BReLU	DPN68+IN+ DC+DRN	21.47	0.7453	0.256
(7)L1+FFT+BReLU	DPN68+IN+ DC+HDMGN	21.51	0.7463	**0.250**
(8)L1+FFT+BReLU	DPN68+IN+DC +DRN+HDMGN	**21.60**	**0.7500**	0.254

Table 2: Ablation study results of loss function and architecture. "DC" denotes that TDN uses deformable convolution for decoder part of EDN and tail convolutional layer instead of standard convolutional layer. "IN" denotes that EDN adds Instance Normalization layers to the skip connections. L1, FFT and BReLU refer to spatial domain loss, frequency domain loss and threshold limitation loss respectively. The lines with the best result are in bold font.

When deformable convolution is used, because 2D offset greatly enhances the ability of CNN transformation modeling, foreground and hazy background region features are well separated to provide different degrees of dehazing.

Not only can deformable convolutional layer help recognizing the density of haze, but also it can fix the artifacts caused by outlier pixels. As shown in Figure 4 right and Figure 3 left, although we do [0,1] clamping to get the output, there are also some extreme outlier pixels that cause the artifacts in the intermediate feature maps and final output. By replacing standard convolutional layers in the decoder and tail convolutional block to deformable convolutional layer, the learnable convolution offsets in the deformable convolutional layer ignore these outliers and choose the suitable pixels to do the convolution operator, which helps fixing these artifacts. The ablation study results as shown in Table 2 setting (4) and setting (5) also demonstrate the effectiveness of deformable convolutional layers.

Figure 4: Left: example output image w/o deformable convolution. Middle: example output image with deformable convolution. Right: example output image w/o deformable convolution that contains artifacts.

Figure 5: Left: hazy image "48.png". Middle: hazy free image of the network removing instance normalization layers from TDN. Right: hazy free image of TDN.

Ablation Studies of Instance Normalization Layer and Proposed Loss Functions. We provide qualitative visual effect comparison of an example image, as shown in Figure 5. After adding instance normalization layers to the skip connections, the color saturation of leaves and grass increases, which getting a better visualized reconstructed hazy free image. Comparison between Table 2 setting (3) and setting (4) demonstrates that results of network with instance normalization layers get better perceptual quality and is more similar to the reference ground truth.

As shown in Figure 6, TDN using all proposed loss functions get the best loss convergence result. Threshold limitation (BReLU) loss helps training the network in both the early stage and the late stage. Adding frequency domain (FFT) loss hinders the training process in the spatial domain

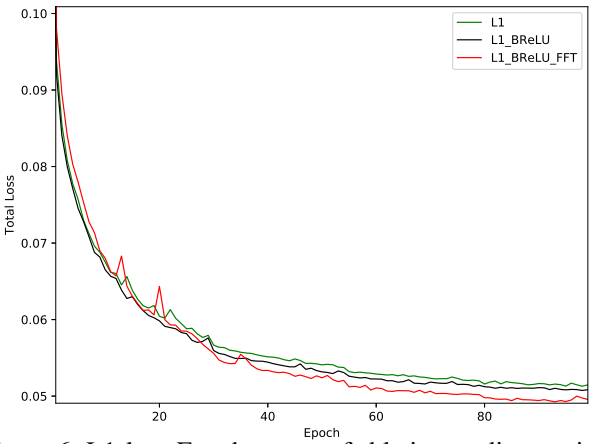

Figure 6: L1_loss-Epoch curves of ablation studies, setting (1) to (3). L1: Only use spatial domain loss as the loss function. L1_BReLU: Use both spatial domain loss and threshold limitation loss as loss functions. L1_BReLU_FFT: Use spatial domain loss, threshold limitation loss and frequency domain loss as loss functions.

in the early stage, but when whole training process finishes, the spatial domain (L1) loss curve converges to a better result, comparing with L1 curve and L1_BReLU curve. Comparing Table 2 setting (1) to (3), we can find that after adding threshold limitation loss, the performance is a little better. Furthermore adding frequency domain loss, regardless of the restoration fidelity to the ground truth (PSNR and SSIM) or the perceptual quality (LPIPS), the reconstructed hazy free image gets much more better performance.

3.3. Comparison with State-of-the-Art Methods

This section illustrates the comparisons between our proposed TDN with the state-of-the-art methods on real-world benchmark datasets I-HAZE, O-HAZE, Dense-Haze and NH-Haze, and synthetic dataset ITS. We re-train the state-of-the-art models of different benchmark on corresponding training sets for fair comparisons. Both quantitative evaluation and visual effect are reported.

Competitors. Six state-of-the-art methods are included in the comparisons: DCP [17], DehazeNet [11], AOD-Net [21], DCPDN [34], GCANet [12] and FFA [25].

Quantitative Results. The quantitative comparison results are shown in Table. 3. On synthesis test-set RESIDE (ITS), our method is second only to FFA. Note that FFA has too many blocks and does all feature mapping operator in the scale of whole image, which results in a huge amount of computation, while our TDN does the main convolution operators on the downsampled scales, our method has a quick inference time. One 1600×1200 image only costs 0.64s to inference. On real-world NTIRE dehazing challenges datasets, our method has the best performance in terms of both PSNR and SSIM, and especially on dense haze and nonhomogeneous haze scene, our method outperforms other methods by a noticeable margin.

Visual effect comparison. As shown in Figure 7 to Figure 11, suffering from color distortion, DCP gets darker results on ITS test set and bluer results on real-world NTIRE dehazing challenges testsets. On NTIRE testsets, most of the color information has been lost in DehazeNet, at the same time, it generates some artifacts. AOD-Net cannot remove haze effectively. On O-Haze testset, DCPDN recovers images whose color is with large deviations. On dense haze test and NH-Haze dataset, DCPDN generates very unpleasant artifacts, FFA gets dirty results, while GCANet over-dehazes the hazy images resulting in very dark wrong results, which demonstrate that DCPDN and GCANet are not robust for dense haze or nonhomogeneous haze scene. Although FFA gets somewhat good results on I-HAZE and O-HAZE testset, but due to the huge graphics memory usage, the inference can only be done with chop and concatenation strategy, which results in serious checkerboard artifacts. Only our proposed TDN reconstructs faithful and sharp hazy free results with little artifact and good perceptual quality on all commonly used dehazing benchmarks.

NTIRE-2020 Dehazing Challenge. For the newly published NTIRE2020-Dehaze dataset, the haze presented in the images are much more nonhomogeneous than normal images in the literature. As shown in Figure 10, the state-of-the-art methods' performances degraded heavily due to the reason that the haze covers the entire image nonhomogeneously with different density in different region. Since TDN can automatically estimate the haze density information from the nonhomogeneous haze image, the dehazed images generated by TDN are much more visually pleasing. We evaluate the quantitative performances of the methods on the NTIRE2020 dataset $41 \sim 45$ since the ground-truth images of validation set is not made available now. As shown in Table 3 (NH-Haze), TDN outperforms all the other state-of-the-art methods.

Table 4 includes the top-6 perceptual quality methods from the contest. It is found that TDN is among the top perceptual quality methods in the NTIRE2020-Dehazing Challenge.

4. Conclusion

In this paper, we propose a novel coarse to fine model Trident Dehazing Network (TDN) with automatic haze density recognition for nonhomogeneous dehazing. Three sub-nets compose TDN, the Encoder-Decoder sub-Net reconstructs the coarse hazy free feature, the Details Refinement sub-Net refines the coarse feature maps with more high frequency details that was lost through pooling layers in the encoder, the Haze Density Map Generation sub-Net can automatically reconstruct the haze density map with no extra supervision. Extensive experimental results demonstrate that TDN is robust on not only synthetic datasets, but also real-world scene with dense haze and nonhomogeneous haze, and outperforms the state-of-the-arts with better fidelity and perceptual.

	RESIDE(ITS)		NTIRE18				NTIRE19 (Dense Haze)		NTIRE20 (NH-Haze)		
			I-Haze		O-Haze						
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	LPIPS
DCP	16.62	0.8179	11.90	0.3842	15.10	0.3546	12.12	0.2467	10.57	0.5181	0.506
DehazeNet	20.65	0.7975	15.93	0.7734	19.99	0.6885	13.84	0.4225	17.00	0.5444	0.695
AOD-NET	19.82	0.8187	16.01	0.7738	18.05	0.6305	13.14	0.4121	15.41	0.5677	0.508
DCPDN	28.16	0.9555	17.43	0.8059	22.51	0.7321	14.48	0.4844	22.74	0.7334	0.285
GCANet	30.07	0.9597	16.50	0.7598	21.86	0.7304	10.71	0.3615	14.27	0.5839	0.396
FFA	**36.37**	**0.9870**	17.20	0.7943	22.74	0.8339	14.39	0.4524	19.87	0.6915	0.295
Ours(Trident)	34.59	0.9754	**19.33**	**0.8287**	**23.90**	**0.7685**	**16.48**	**0.5490**	**23.06**	**0.7554**	**0.250**

Table 3: The PSNR/SSIM/LPIPS of different methods over SOTS-indoor, I-HAZE, O-HAZE, Dense Haze and NH-Haze testset. The lines with the best result are in bold font.

Team	Contest Method	PSNR↑	SSIM↑	LPIPS↓	MOS↓ Ranking
ours	TDN	21.41	**0.71**	**0.267**	**1**
other teams	method1	20.85	0.69	0.285	2
	method2	21.60	0.67	0.363	3
	method3	**21.91**	0.69	0.361	4
	method4	20.11	0.66	0.351	5
	method5	19.70	0.68	0.301	6

Table 4: The average PSNR/SSIM/LPIPS/MOS ranking of top perceptual quality methods over NTIRE2020 test dataset. The lines with the best result are in bold font.

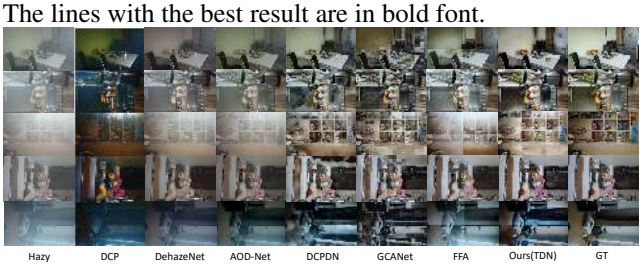

Figure 7: The visual results of NTIRE2018 I-HAZE validation dataset.

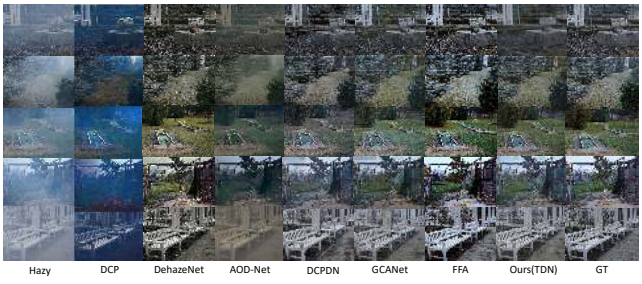

Figure 8: The visual results of NTIRE2018 O-HAZE validation dataset.

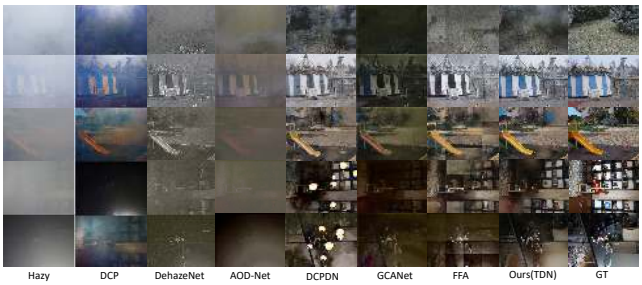

Figure 9: The visual results of NTIRE2019 Dense Haze validation dataset.

Figure 10: The visual results of NTIRE2020 NH-Haze dataset $41 \sim 45$.

Figure 11: Example qualitative comparisons on SOTS (indoor) testset.

5. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61972157, 61902129, 61772524, 61876161, 61701235, 61373077, 61602482), the National key technologies R&D program of China (SQ2019YFC150159), the Science and Technology Commission of Shanghai Municipality Program (18D1205903), the Shanghai Pujiang Talent Program (19PJ1403100), the Science and Technology Commission of Pudong (PKJ2018Y46), the Beijing Municipal Natural Science Foundation (4182067); and partly by the Fundamental Research Funds for the Central Universities associated with Shanghai Key Laboratory of Trustworthy Computing.

References

[1] Cosmin Ancuti, Codruta O Ancuti, and Radu Timofte. Ntire 2018 challenge on image dehazing: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 891–901, 2018.

[2] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte et al. Ntire 2019 challenge on image dehazing: Methods and results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[3] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In *arXiv:1904.02904*, 2019.

[4] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. NH-HAZE: An image dehazing benchmark with nonhomogeneous hazy and haze-free images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[5] C. O. Ancuti, C. Ancuti, and R. Timofte. NH-HAZE: An image dehazing benchmark with nonhomogeneous hazy and haze-free images. *IEEE CVPR, NTIRE Workshop*, 2020.

[6] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[7] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: a dehazing benchmark with real hazy and haze-free indoor images. In *arXiv:1804.05091v1*, 2018.

[8] C. O. Ancuti, C. Ancuti, Florin-Alexandru Vasluianu, and R. Timofte et al. Ntire 2020 challenge on nonhomogeneous dehazing. *IEEE CVPR, NTIRE Workshop*, 2020.

[9] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. Ntire 2020 challenge on nonhomogeneous dehazing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

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

[11] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[12] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1375–1383. IEEE, 2019.

[13] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in neural information processing systems*, pages 4467–4475, 2017.

[14] Raanan Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):1–9, 2008.

[15] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014.

[16] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[17] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[20] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep photo: Model-based photograph enhancement and viewing. *ACM transactions on graphics (TOG)*, 27(5):1–10, 2008.

[21] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4770–4778, 2017.

[22] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018.

[23] Zhuwen Li, Ping Tan, Robby T Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4988–4997, 2015.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[25] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. *arXiv preprint arXiv:1911.07559*, 2019.

[26] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pages 154–169. Springer, 2016.

[27] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018.

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016.

[29] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2995–3000, 2014.

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[31] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[33] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018.

[34] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018.

[35] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Joint transmission map estimation and dehazing using deep networks. *arXiv preprint arXiv:1708.00581*, 2017.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[37] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing*, 24(11):3522–3533, 2015.

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.