Distilling Image Dehazing with Heterogeneous Task Imitation

Ming Hong[1] Yuan Xie[2*] Cuihua Li[1] Yanyun Qu[1†]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Informatics, Xiamen University, Fujian, China
[2]School of Computer Science and Technolog,
East China Normal University, Shanghai, China
mingh@stu.xmu.edu.cn, yxie@cs.ecnu.edu.cn, chli@xmu.edu.cn, yyqu@xmu.edu.cn

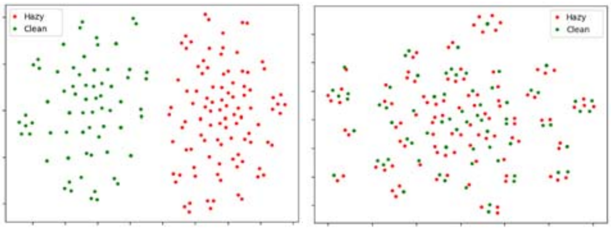
Abstract

State-of-the-art deep dehazing models are often difficult in training. Knowledge distillation paves a way to train a student network assisted by a teacher network. However, most knowledge distill methods are used for image classification and segmentation as well as object detection, and few investigate distilling image restoration and use different task for knowledge transfer. In this paper, we propose a knowledge-distill dehazing network which distills image dehazing with the heterogeneous task imitation. In our network, the teacher is an off-the-shelf auto-encoder network and is used for image reconstruction. The dehazing network is trained assisted by the teacher network with the process-oriented learning mechanism. The student network imitates the task of image reconstruction in the teacher network. Moreover, we design a spatial-weighted channel-attention residual block for the student image dehazing network to adaptively learn the content-aware channel level attention and pay more attention to the features for dense hazy regions reconstruction. To evaluate the effectiveness of the proposed method, we compare our method with several state-of-the-art methods on two synthetic and real-world datasets, as well as real hazy images.

1. Introduction

The performance of the high-level tasks in the computer vision depends on the quality of the input image. However, images captured in hazy scenes always suffer from degradation of color and details. To remove haze and improve visibility of the hazy image, many dehazing methods [8, 36, 28, 5] have been proposed.

Directly mapping a hazy image to its corresponding clean target is difficult due to a lot of information lost. Many previous methods focused on designing the dehazing networks with the atmospheric scattering model [24, 15], but the inaccurate estimation of parameters may produce a cumulative error and degrade the quality of the dehazing results [18]. Hence, it's important to study how to improve the performance of end-to-end dehazing networks.



(a) (b)

Figure 1. Intermediate features visualized by t-SNE [22]. (a) Without the supervision of Teacher. (b) With the supervision of Teacher. ● denotes features extracted by a dehazing network with the hazy input. ● denotes features extracted by Teacher with the clean input. The proposed method aims to make the features extracted from the hazy image similar to those extracted from the clean image.

Inspired with the success of knowledge distillation [10, 30, 29] in image classification [26], image segmentation [21], and object detection [29], we utilize knowledge distillation to assist in training the dehazing model. The original knowledge distillation is realized through the teacher-student learning paradigm and the student network (Student) with a light-weight model learns from the teacher network (Teacher) which is a pre-trained heavy-weight model. In the early applications, knowledge distillation is used to compress the student network. Recently, it is implemented on knowledge transfer between two deep models.

Until now, the knowledge distillation is mainly applied in image classification, image segmentation, and object detection, while it is hardly implemented on image restoration. This paper investigates how to conduct knowledge distillation on image dehazing. There exist three difficulties to

[*] Equal contribution
[†] Corresponding author

distill image dehazing:

1) Which kind of network for what task can give effect knowledge to assist the dehazing network in training? Most existing knowledge distillation methods mainly use Teacher and Student to deal with the same task, for example, both Teacher and Student are used for image classification or segmentation et al. Few investigate knowledge transfer between two networks with different tasks. Could the heterogeneous task help to train the deep dehazing model?

2) How does Teacher assists in training Student? Most knowledge distillation methods focus on result-oriented learning while neglecting the process-oriented learning. For example, Hinton [10] only measures the similarity of class distribution between the outputs of Teacher and Student to control the two network learning and neglect the rich dark knowledge in the middle learning process. How does the dark knowledge in the middle learning process be utilized to help training Student?

3) How is the similarity between Teacher and Student measured? In knowledge distillation of image classification, the class distribution is used to measure the similarity between the teacher and student networks. However, as for image restoration, none of the class distributions can be used.

To solve the problems, we propose the knowledge distilling dehazing network (KDDN) which distills image dehazing with the heterogeneous task network. To solve the first problem, we use the network for image reconstruction to assist the dehazing network in training. The off-the-shelf auto-encoder network is treated as Teacher. Moreover, to solve the second problem, we adopt the process-oriented learning mechanism for knowledge transfer. We supervise the intermediate features and use the feature similarity to control the imitation learning from the image reconstruction of Teacher to image dehazing of Student. The process-oriented supervision makes full use of the dark knowledge for better image dehazing results. For the solution of the third problem, we use the image fidelity loss function, the perceptual loss function and the difference of the feature maps generated in the pair-wise intermediate process between Teacher and Student. In Figure 1, the immediate features learned from the dehazing network with and without the supervision of Teacher are compared. With the supervision of Teacher, the features of the dehazing network are similar to those of Teacher.

Moreover, considering the uneven concentration of haze in an image and yield more efficient representations for single image dehazing, we propose a haze density-aware imitation loss to upweight the dense hazy regions. We design spatial-weighted channel attention residual block for the student network, which adaptively recalibrates the important features for image dehazing.

In summary, the main contributions of this work are as follows:

- We propose a knowledge distillation method for image dehazing with the heterogeneous task. Teacher is an off-the-shelf auto-encoder network for image reconstruction, which assists the image dehazing network in training. To our best knowledge, this is the first work to implement knowledge distillation on image restoration with a heterogeneous task.
- The process-oriented learning mechanism is proposed. The process-oriented loss function, the frequently used result-oriented fidelity loss and perceptual loss are combined to form the total loss to measure the similarity of features between Teacher and Student.
- We design the spatial-weighted channel-attention residual block and a haze density aware imitation loss for the dehazing network.

2. Related Work

2.1. Single Image Haze Removal

The existing dehazing methods are divided into two classes: the physical-model dependent methods and the model-free methods. The traditional model-based methods design many priors to estimate atmospheric light and transmission map, then recover the clear image based on the atmospheric scattering model, such as dark channel prior [8], color attenuation prior [8] and non-local prior [5].

With the rising of deep learning, many deep dehazing models depending on the atmospheric scattering model spring up. In [24], a coarse-to-fine multi-scale convolutional neural network (MS-CNN) was proposed to estimate the transmission map. In [6], an end-to-end network with a bilateral rectified linear unit was proposed to estimate the transmission map. Li *et al.* [15] re-formulated the atmospheric scattering model and jointly estimated transmittance and atmospheric light. Furthermore, Zhang *et al.* [33] directly embedded the physical model into the dehazing network via a mathematical operation module and proposed a densely connected encoder-decoder structure to leverage features. Yang *et al.* [32] introduced a disentanglement and reconstruction network independent of the physical model to generate realistic haze-free images trained by an adversarial process with unpaired data.

As the inaccurate estimation of intermediate results may produce a cumulative error and degrade the quality of the dehazing results, model-free deep dehazing methods that directly learn the map between hazy input and clean result attract more and more attention. Liu *et al* [20] designed a dual residual network to exploit the potential of paired operations for image restoration tasks. Li *et al.* [17] proposed a haze removal network trained with strong supervision and enhanced the detail and color of dehazing result

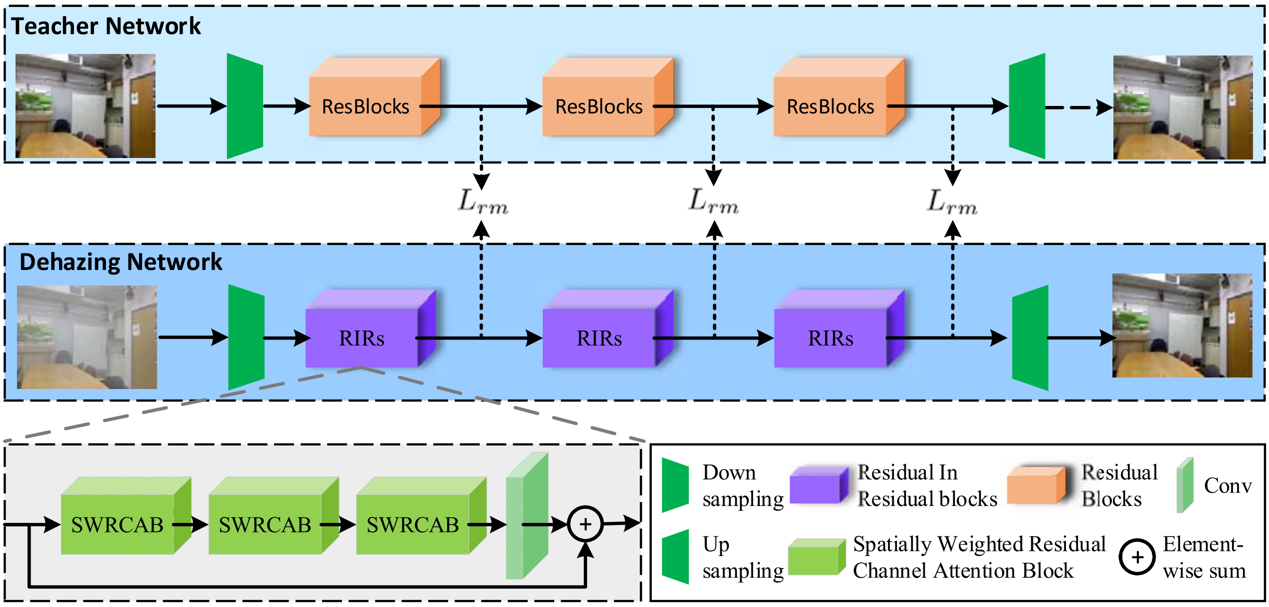


Figure 2. An overview of the knowledge distilling dehazing network.

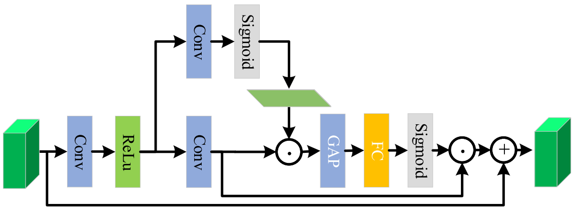


Figure 3. The structure of spatially weighted residual channel attention block (SWRCAB).

with a refinement network trained with weak supervision. Qu *et al.* [23] proposed an enhanced Pix2pix dehazing network (EPDN) to get perceptually pleasing images. Liu *et al.* [19] introduced an attention-based multi-scale estimation network named GridDehazeNet for image dehazing. These methods only calculate the difference between the output dehazing image and ground truth clean image for supervised learning. Different from them, the proposed method attempts to learn more efficient intermediate representation by forcing the dehazing network directly to imitate the feature extracted from the clean image domain.

2.2. Network and Feature Mimicking

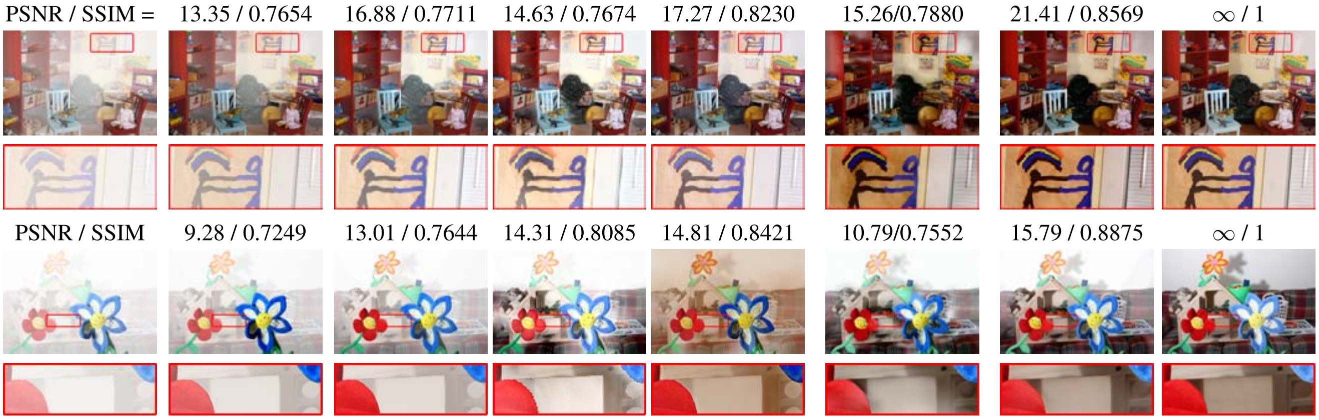
Network mimicking is originally introduced to distill knowledge from Teacher to Student to obtain small and fast models [10] by approximating the soft output of Teacher. [26] extended this idea to feature mimicking which aims to train a deeper and thinner network by imitating not only the output but also the intermediate representation learned by Teacher. Recently, feature mimicking are widely applied to person re-identification [30], semantic segmentation [21], and object detection [29]. In these approaches, the knowledge is distilled between the common task. Differently, in [11], Hou *et al* extended the idea of feature mimicking to transfer knowledge between large networks that perform heterogeneous tasks. In our work, we investigate the possibility of representation mimicking that makes the intermediate features of Student similar to those of the clean image in the feature distribution to yield more effect features and better dehazing performance. Different from earlier works that have the same input, the inputs of Teacher and Student are different in this paper. Specifically, Teacher is used to extract features from a clean image that is the mimicking objective of Student.

3. Proposed Method

The overall architecture of KDDN consists of Teacher and Student / the dehaizng network, is shown in Figure 2. Specifically, Teacher aims to provide intermediate feature representation of clean images, and Student aims to restore a clean image from its hazy counterpart by transforming the intermediate features into the clean image domain.

3.1. Network Architecture

Teacher / Autoencoder. We adopt the encoder-decoder architecture proposed in [14] as Teacher which consists of a down sampling module, a backbone module, and an upsampling module. The down sampling module quarters the input image in terms of side length by using two 3×3 convolution layers with the stride 2 following a ReLU per layer. The backbone module gives different representations of



(a) Input image (b) AOD-Net [15] ICCV'17 (c) PDN [31] ECCV'18 (d) GFN [25] CVPR'18 (e) EPDN [23] CVPR'19 (f) GridDehaze [19] ICCV'19 (g) Ours (h) Ground Truth

Figure 4. The qualitative results of the state-of-the-art methods on Middlebury [7].

the input image from a low level to the high level, here we use 6 residual blocks to construct it. Note that in this paper the residual blocks are built in the form of Convolution-ReLU-Convolution-Add without Batch normalization. The up-sampling module upscales the features to the original size using two 3×3 convolution layers followed by ReLU and bilinear upsampling layers. Then we can obtain the reconstruction result utilizing a 3×3 convolution layer followed by a Tanh activation function. Note that all convolution layers have 64 channels except the input and output layers in this paper.

Student / Dehazing Network. The dehazing network has a similar architecture as Teacher except for the backbone module. Inspired by [35] which employed short and long skip connection to form a very deep network, the backbone in the dehazing network consists of several Residual In Residual (RIR) blocks, as shown in Fig. 2. RIR is generally made up of serval residual blocks (ResBlock) [9] or squeeze and excitation residual blocks (SEResBlock) [12]. However, SEResBlock employs a global average pool operation to learn the weight of each channel that equally aggregates all input features, ignoring the inconsistent concentration of haze. To pay more attention to seriously degraded regions and informative channels, we introduce a Spatially Weighted Residual Channel Attention Block (SWRCAB) to learn the content-aware channel level attention, which is depicted in Figure 3. Different from SEResBlock, SWRCAB first learns spatially weights of input features using a convolutional layers followed by a Sigmoid layer, then it obtains the spatially weighted features via the element-wise multiplication, and it gets each channel's attention via a global average pooling layer which is followed by a linear transform layer and a Sigmoid activation layer. Our KDDN contains 6 RIRs each of which contains 3 SWRCABs.

3.2. Loss for Teacher Network

To achieve an effective feature representation from Teacher, we train it in an unsupervised style. Then the objective function of Teacher is formulated as:

$$L_T = \|J - T(J)\|_1 , \tag{1}$$

where T is the transform function of Teacher, and J is the input clean image.

3.3. Loss for Dehazing Network

To achieve the goal of domain transfer from Teacher to the dehazing network and mimic features of Teacher, we formulate the overall loss function:

$$L_S = L_r + \lambda_p L_p + \lambda_{rm} L_{rm}, \tag{2}$$

where L_r denotes the reconstruction loss, L_p denotes the perceptual loss, and L_{rm} means the representation mimicking loss that aims to mimic the representations of clean image domain. λ_p and λ_{rm} are the balancing weights.

Reconstruction Loss. To obtain the reconstruction result, we adopt the mean absolute error (MAE) to measure the difference between the ground-truth and dehazing result:

$$L_r = \|S(I) - J\|_1 , \tag{3}$$

where $S(I)$ denotes the dehazing result, and J denotes the ground-truth.

Perceptual Loss. Perceptual loss aims to measure the global discrepancy between dehazing result and corresponding ground-truth in terms of the deep features, L_p is expressed as:

$$L_p = \sum_{i=1}^{N} \frac{1}{M_i} \left\| \Phi^i(S(I)) - \Phi^i(J) \right\|_1 , \tag{4}$$

where Φ^i denotes the operator of feature extraction of i^{th} layer with M_i elements of a pre-trained deep network. In our method, we compute the feature loss at layers 2, 7, 12, 21, and 30 of VGG19 trained on ImageNet for image classification.

Representation Mimicking Loss. As the intermediate representations of the clean image extracted by Teacher have abundant knowledge for reconstruction. Mimicking the feature distribution of the clean images might be helpful for training the dehazing network. Hence, we introduce two different types of representation mimicking loss: feature matching imitation loss (FMIL) and haze density aware imitation losses which weights FMIL (WFMIL). For simplicity, we let $T^m(J)$ denotes the feature maps of the m^{th} layer of Teacher trained on clean images, let $S^n(I)$ denotes the feature maps of the n^{th} layer of Student trained on hazy images, and let g be a linear transformation function such as a pointwise convolution which aims to make sure the number of channels of $T^m(J)$ and $S^n(I)$ is consistent.

FMIL is formulated as:

$$L_{rm} = \sum_{(m,n)\in C} \left| T^m(J) - g\left(S^n(I)\right) \right|, \tag{5}$$

where C is a set of candidate pairs of feature representations.

Since dense hazy regions are seriously degraded caused by the uneven concentration of haze in an image, the dehazing network often fails in dehazing well in the earlier stage. To overcome this issue, we design a haze density aware imitation loss to up-weight the dense hazy region. Specifically, we propose to use the concentration of haze to weight the representation mimicking loss. In our solution, we employ a transmission map and the residual between the hazy image and ground-truth in gray to measure the concentration of haze. As shown in Figure 5 (c), the transmission map is highly related to the scene depth and provides smooth guidance because it ignores the details of objects. However, the transmission maps are not always available in the training set. Hence, we measure the concentration of haze by calculating the residual between the hazy image and ground-truth in gray, as shown in Figure 5 (d). With these two measurements, WFMIL is formulated as:

$$\begin{aligned} L_{wrm} &= \sum_{(m,n)\in C} \Psi * \left| T^m(J) - g\left(S^n(I)\right) \right|, \\ \Psi_1 &= norm(1-t), \\ \Psi_2 &= norm\left(\left|I_g - J_g\right|\right), \end{aligned} \tag{6}$$

where t is the transmission map obtained from the training set, I_g and J_g are the gray image of I and J, respectively. $norm()$ denotes a normalization operator which makes the pixel values normalized to $[0,1]$. Specifically, for a given image I,

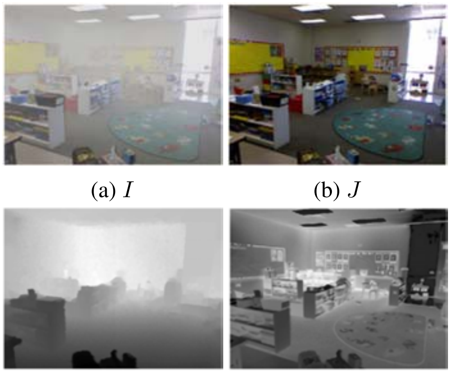

(a) I (b) J
(c) $norm(1-t)$ (d) $norm\left(\left|I_g - J_g\right|\right)$

Figure 5. An example of the concentration of haze measurement.

$$norm(I(i,j)) = (I(i,j) - min(I))/(max(I) - min(I)), \tag{7}$$

where $min(I)$ denotes the minimum value of I, $max(I)$ denotes the maximum value of I, (i,j) is the pixel location.

4. Experiments

4.1. Experiment Setup

Dataset. In this paper, we evaluate the proposed method on the indoor training dataset (**ITS**) of RESIDE [16]. ITS contains 13, 990 synthetic hazy images generated from two indoor scenes NYU Depth V2. We use the indoor dataset **SOTS** of RESIED and **Middlebury** used in D-HAZE [7] for testing. SOTS consists of 500 indoor hazy images generated from 50 different scenes of NYU Depth V2 and Middlebury contains 23 hazy images generated from high quality real scenes.

We also evaluate the proposed method on a real-world dataset **O-HAZE** [3], which contains 45 pairs of hazy images and the corresponding haze-free images of various outdoor scenes. Especially, the hazy images in this dataset are shut in the real haze scene. For the fairness of comparison, we follow the setup of the NTIRE 2018 Image Dehazing Challenge [1] which uses the first 35 pairs for training, the remaining 5 pairs for validation and 5 pairs for testing..

Implementation Details. We use PyTorch 1.2.0 to implement our models with an NVIDIA RTX 2080 GPU. Teacher and Student are trained separately. We first train Teacher on the clean images for 30 epochs, and then fix it and train the dehazing network for another 60 epochs. KDDN are trained by the Adam optimizer with the initial learning rate of 10^{-4} and a momentum of 0.9. The cyclincal cosine learning rate schedule technique [4] is used to adjust the learning rate which is gradually decayed every 20 epochs. The weights λ_p and λ_{rm} are empirically set as 1. m is set to 2, 4, and 6, respectively, and n is set to 1, 2, and 3 respectively. PSNR

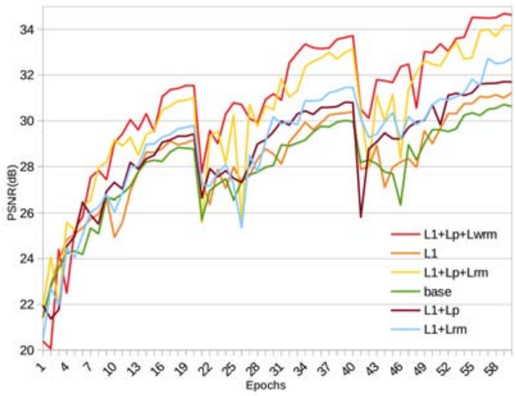


Figure 6. The PSNR values of different settings in 60 epochs.

and SSIM are used as the criteria to evaluate the dehazing performance.

4.2. Ablation Study

We conduct the ablation study to investigate the loss functions, the effect of process-oriented feature mimicking, the effect of SWRCAB, and the sensitivity of Teacher. The baseline dehazing network contains 6 RIR blocks constructed by SEResBlock [13] and trained with L1 loss, denoted as "**Baseline**". We construct six variants with SWRCAB and different loss functions.

(1) L_1: SEResBlock is replaced with SWRCAB to construct the dehazing network and train with L1 reconstruction loss.

(2) L_1 + L_{ds}: Similar to the architecture of the dehazing network, the overall loss functions contains two terms: the L1 reconstruction loss and the deep supervision loss L_{ds}. Deep supervision means that not only the final output but also the intermediate layers are supervised by the ground truth. In our experiment, the supervision is implemented on the same layers of KDDN.

(2) L_1+L_p: The sum of the L1 reconstruction loss and the perceptual loss is used to train the dehazing network.

(3) L_1+L_{rm}: The sum of the L1 reconstruction loss and FMIL is used to train the dehazing network.

(4) L_1+L_p+L_{rm}: The sum of the L1 reconstruction loss, the perceptual loss and FMIL is used to train the dehazing network.

(5) L_1+L_p+L_{wrm}: The sum of the L1 reconstruction loss, the perceptual loss and WFMIL is used to train the dehazing network.

The average PSNR and SSIM values of six variants are reported in Table 1. Obviously, the dehazing network with FMIL achieves better dehazing results than Baseline. Specifically, we can observe the following results: (1) SWRCAB can improve dehazing performance more effectively than SEResBlock and the gain is (0.71, 0.0031) in terms of PSNR and SSIM. (2) Both traditional perceptual loss and FMIL improve the dehazing performance, and FMIL performs better and the gain is (1.06, 0.0046). (3) The proposed loss functions, L1+Lp+Lrm and L1+Lp+Lwrm, perform best under the supervision of ground-truth image and intermediate features. (4) WFMIL achieve better dehazing performance than FMIL and the gain is (0.57, 0.0007).

Methods	PSNR	SSIM
Baseline	30.52	0.9613
L_1	31.23	0.9644
L_1+L_{ds}	31.86	0.9749
L_1+L_p	31.69	0.9726
L_1+L_{rm}	32.75	0.9772
L_1+L_p+L_{rm}	34.15	0.9838
L_1+L_p+L_{wrm}	34.72	0.9845

Table 1. Ablation study on the effect of loss functions.

To analysis the effect of the optimization process of the dehazing network, Figure 6 presents the PSNR performance with 60 training epochs under different settings. It indicates that the proposed method converges faster and obtain a better result.

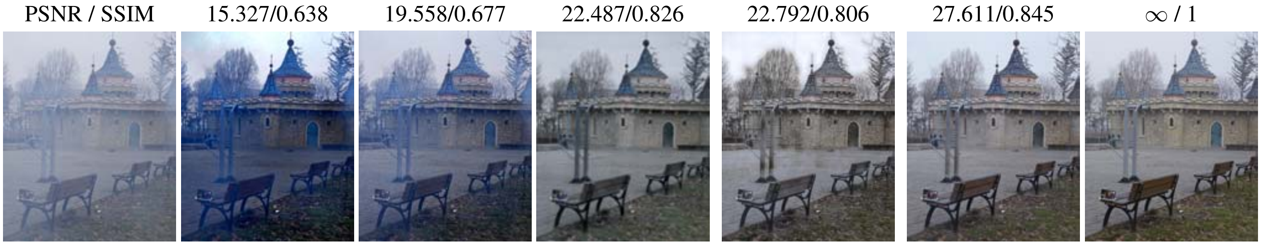
Process-oriented Learning vs Result-oriented Learning. KDDN is viewed as a process-oriented approach, while the state-of-the-art dehazing methods are result-oriented. Specifically, the loss measures the representation difference between dehazing result and the clean target image, while KDDN measures the difference of intermediate features between the dehazing network and Teacher. (2) KDDN employs an autoencoder network for image reconstruction to assist the dehazing network in training. While the state-of-the-art methods contain only a streamline for image dehazing. KDDN can achieve a superior result, as shown in Table 1 and Figure 6.

Network	Metrics	Random	w/o	wM	wH
Teacher	PSNR	6.61	-	33.47	43.04
	SSIM	0.033	-	0.9634	0.9902
Student	PSNR	27.52	31.23	31.82	32.75
	SSIM	0.9475	0.9644	0.9708	0.9772

Table 2. Effects of Teacher on the dehazing network. '-' denotes the results without using a teacher network.

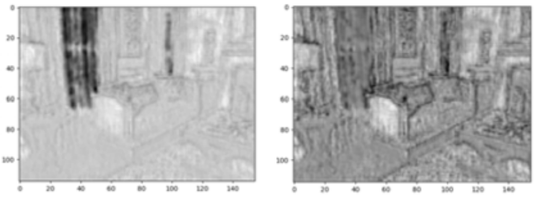
Effectiveness of Representation Mimicking. Our model aims to make the distribution of features of the dehazing network similar to that of Teacher. We randomly select 100 images from SOTS indoor dataset and plot the features of the 4th RIR of the dehazing network and the 2nd ResBlock of Teacher using t-SNE [22]. As shown in Figure 1 (b), with the proposed representation mimicking loss we can achieve better feature representation. The quantitative results given in Table 1 also indicate that the proposed method can obtain a superior PSNR and SSIM performance than the baseline model trained with only L1 loss.

Effectiveness of SWRCAB. The features before and after



(a) Input image (b) DCP [8] (c) GFN [25] (d)[34] (e) GridDehaze [19] (f) Ours (g) GT

Figure 7. Visual Comparison of images from O-HAZE for image dehazing.



(a) Before spatially weighting (b) After spatially weighting

Figure 8. Illustration of the features before and after the proposed spatially weighting module.

spatially weighting are visualized in Figure 8 (a) and (b). The proposed spatially weighting can refine the feature and pay different attention to spatial positions. The quantitative results given in Table 1 also verify the effectiveness of SWRCAB.

Sensitivity of Teacher. The proposed method employs an autoencoder network to extract the intermediate representations of the clean images. To investigate the effects of Teacher on the performance of the dehazing network, we conduct four different experiments using L1 loss with different settings of Teacher: Randomly initialized Teacher (Random), without the supervision of Teacher (w/o), with the supervision of a mid performance Teacher (wM), with the supervision of a high performance Teacher(wH). It demonstrates that the bad Teacher is unhelpful to optimize the dehazing network while the good Teacher improve the dehazing performance.

Method	SOTS [16]		Middlebury [7]	
	PSNR	SSIM	PSNR	SSIM
AOD-Net [15]	19.06	0.8504	13.40	0.7979
PDN [31]	17.33	0.8164	14.64	0.8088
GFN [25]	22.30	0.8800	13.27	0.7514
EPDN [23]	25.06	0.9232	15.28	0.8096
GridDehaze [19]	32.16	0.9836	14.21	0.7783
KDDN (Ours)	34.72	0.9845	17.27	0.8676

Table 3. Comparisons of six methods on two synthetic datasets in terms of PSNR and SSIM. We highlight the best results in red color and the second-best results in blue color.

4.3. Results on Synthetic Datasets

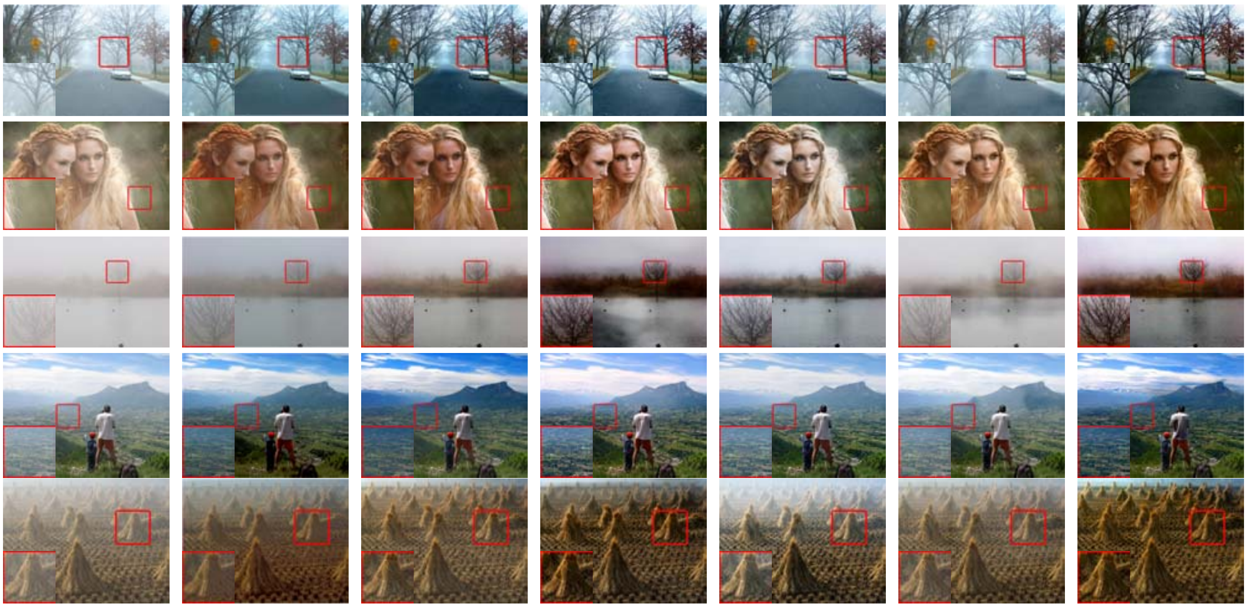
We compare KDDN with five state-of-the-art methods: PDN [31], AOD-Net [15], GFN [25], Grid- Dehaze [19], and EPDN [23] on SOTS and Middlebury [7]. The average PSNR and SSIM values are reported in Table 3. Note that all models are retrained on ITS using the codes provided by the authors and all models have not been retrained or fine-tuned on Middlebury. It is obvious that KDDN outperforms other state-of-the-arts with the gain of at least 2.56dB/0.0009 in terms of PSNR and SSIM. KDDN trained on ITS not only achieves significant improvement on SOTS but also is generalized well on Middlebury.

Figure 4 give the comparison of KDDN and the state-of-the-art methods in visual effects on two synthetic images from Middlebury. It is observed that PDN, AOD-Net and GFN tend to underestimate the haze level and the ouputs still look fairly hazy. GridDehaze, on the other hand, still has some remaining haze. Although EPDN achieves superior performance, it suffers from color distortion. In contrast, the dehazing results of KDDN are more close to the groundtruth and can generate better details, which are also verified by the PSNR and SSIM value.

4.4. Results on Real-world Hazy Scenes

Results on O-HAZE. We follow the setting of the NTIRE 2018 Image Dehazing Challenge [1] to evaluate the proposed method. For the fairness of comparison, we obtain the results of GridDehaze [19] by re-training them on O-HAZE. Table 4 presents the top 5 results of O-HAZE Challenge and the comparison results with GPN and GridDehaze. Obviously, KDDN achieves the best results in terms of PSNR and SSIM, and outperforms the ranking first team by 0.857dB/0.003. It demonstrates that KDDN can perform well on real-world scenes, the visual comparison presented in Figure 7 also verifies it. We also present the dehazing result of a heavy hazy image from DENSE-HAZE dataset [2] in Figure 10. It shows that the dehazing result has less haze remaining and generates vivild colors.

Results on Real Hazy Photographs. Figure 9 presents the qualitative comparison of real hazy photographs collected by previous work. As shown in Figure 9, DCPDN and

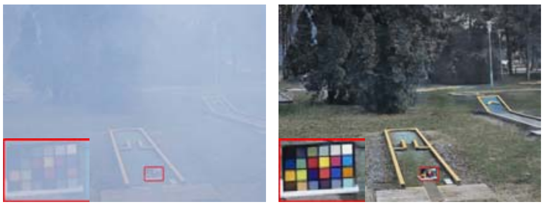


(a) Hazy input (b) AOD-Net [15] (c) PDN [31] (d) GFN [25] (e) EPDN [23] (f) GridDehaze [19] (g) Our method

Figure 9. Visual comparisons of real-world hazy photographs. Our method has less haze remaining and generates images with vivid colors.

method	PSNR	SSIM
BJTU	24.598	0.777
KAIST-VICLAB [27]	24.232	0.687
Scarlet Knights [34]	24.029	0.775
FKS	23.877	0.775
Dq-hisfriends	23.207	0.770
GFN [25]	17.645	0.612
GridDehaze [19]	21.913	0.730
KDDN (L1+Lp+Lrm)	25.312	0.775
KDDN (L1+Lp+Lwrm)	25.455	0.780

Table 4. Comparison on O-HAZE dataset in terms of PSNR and SSIM. We highlight the best results in red color and the second-best results in blue color.



(a) Hazy input (b) Our method

Figure 10. Visual of a dehazing result on dense a hazy image.

AOD-Net tend to remain haze, while EPDN and GFN suffer from color distortions. Especially, GridDehaze performs poor on real hazy images, which indicates it overfits on the training set. Contrarily, the proposed method is capable of generating more visual pleasant results than other methods.

5. Conclusion

In this paper, we propose a knowledge distillation method for image dehazing in which Student imitates the heterogeneous task of Teacher. Teacher is an off-the-shelf auto-encoder network for image reconstruction. The process-oriented mechanism is adopted to train the teacher and student network. Moreover, a spatial-weighted channel-attention residual block (SWRCAB) is designed for the image dehazing network, which adaptively pays more attention to the important features for dense hazy regions reconstruction. We conduct extensive experiments on two synthetic datasets and real-world datasets, as well as real hazy images. The experimental results demonstrate the proposed method outperform the SOTAs dehazing methods quantitatively and qualitatively.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61876161, Grant 61772524, Grant U1065252 and partly by the Beijing Municipal Natural Science Foundation under Grant 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

CVPR Workshops, June 2018. 5, 7

[2] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, Luc Van Gool, Lei Zhang, and Ming-Hsuan Yang. Ntire 2019 image dehazing challenge report. CVPR Workshops, 2019. 7

[3] 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 *CVPR, NTIRE Workshop*, 2018. 5

[4] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 5

[5] Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016. 1, 2

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Processing*, 25(11):5187–5198, 2016. 2

[7] Christophe De Vleeschouwer Cosmin Ancuti, Codruta O. Ancuti. D-hazy: A dataset to evaluate quantitatively dehazing algorithms. In *IEEE International Conference on Image Processing (ICIP)*, ICIP'16, 2016. 4, 5, 7

[8] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1956–1963, 2009. 1, 2, 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3

[11] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning to steer by mimicking features from heterogeneous auxiliary networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8433–8440, 2019. 3

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 4

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 6

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, July 2017. 3

[15] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017. 1, 2, 4, 7, 8

[16] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Processing*, 28(1):492–505, 2019. 5, 7

[17] Chongyi Li, Jichang Guo, Fatih Porikli, Chunle Guo, Huazhu Fu, and Xi Li. Dr-net: Transmission steered single image dehazing network with weakly supervised refinement. *CoRR*, abs/1712.00621, 2017. 2

[18] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, 2019. 1

[19] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, October 2019. 3, 4, 7, 8

[20] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *CVPR*, June 2019. 2

[21] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, June 2019. 1, 3

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 1, 6

[23] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *CVPR*, June 2019. 3, 4, 7, 8

[24] Wenqi Ren, Si Liu, Hua Zhang, Jin-shan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 154–169, 2016. 1, 2

[25] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, pages 3253–3261, 2018. 4, 7, 8

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 3

[27] Hyeonjun Sim, Sehwan Ki, Jae-Seok Choi, Soomin Seo, Saehun Kim, and Munchurl Kim. High-resolution image dehazing with respect to training losses and receptive field sizes. In *CVPR Workshops*, June 2018. 8

[28] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *CVPR*, pages 2995–3000, 2014. 1

[29] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, June 2019. 1, 3

[30] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *CVPR*, June 2019. 1, 3

[31] Dong Yang and Jian Sun. Proximal dehaze-net: a prior learning-based deep network for single image dehazing. In *ECCV*, pages 702–717, 2018. 4, 7, 8

[32] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[33] He Zhang and Vishal M. Patel. Densely connected pyramid dehazing network. In *CVPR*, pages 3194–3203, 2018. 2

[34] He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Multi-scale single image dehazing using perceptual pyramid deep network. In *CVPR Workshops*, June 2018. 7, 8

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 4

[36] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Processing*, 24(11):3522–3533, 2015. 1