LAP-Net: Level-Aware Progressive Network for Image Dehazing

Yunan Li[1,2] Qiguang Miao[1,2] * Wanli Ouyang[3] Zhenxin Ma[1,2]
Huijuan Fang[1,2] Chao Dong[4] Yining Quan[1,2]
[1] School of Computer Science and Technology, Xidian Univeristy, China
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, China
[3] The University of Sydney, SenseTime Computer Vision Research Group, Australia
[4] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
yn_li@stu.xidian.edu.cn, {qgmiao, ynquan}@xidian.edu.cn, wanli.ouyang@sydney.edu.au,
longmao54sh@hotmail.com, hjfang95@163.com, chao.dong@siat.ac.cn

Abstract

In this paper, we propose a level-aware progressive network (LAP-Net) for single image dehazing. Unlike previous multi-stage algorithms that generally learn in a coarse-to-fine fashion, each stage of LAP-Net learns different levels of haze with different supervision. Then the network can progressively learn the gradually aggravating haze. With this design, each stage can focus on a region with specific haze level and restore clear details. To effectively fuse the results of varying haze levels at different stages, we develop an adaptive integration strategy to yield the final dehazed image. This strategy is achieved by a hierarchical integration scheme, which is in cooperation with the memory network and the domain knowledge of dehazing to highlight the best-restored regions of each stage. Extensive experiments on both real-world images and two dehazing benchmarks validate the effectiveness of our proposed method.

1. Introduction

Outdoor images often suffer from the inclement weathers, such as fog and haze, leading to the degradation of color and textures for distant objects. The problem caused by fog and haze has become a critical issue in many applications like visual surveillance, remote sensing, and intelligent transportation. Plenty of techniques are proposed for dehazing [26], and most of them are based on the atmospheric scattering model [29]. Extra information [19, 32] and multiple images [37, 31, 10] are solutions in the early stage. Then single image dehazing techniques [14, 54, 4] gradually become popular because of their efficiency. The powerful feature representation ability of deep learning also promotes the application of neural networks in haze removal [34, 20, 35, 48].

*Corresponding author

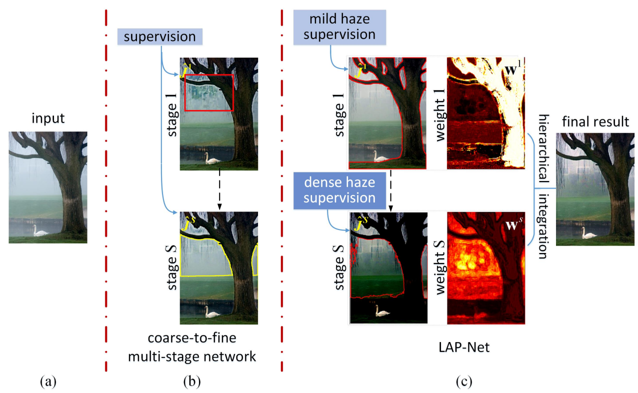

Figure 1. The principle of LAP-Net and comparison with a general multi-stage scheme in a coarse-to-fine fashion. In (b), different stages in the coarse-to-fine network have the same supervision and can only refine the details of stage 1 as marked in the red box. It cannot handle the condition where haze varies in close and distant regions even in the last stage as marked in yellow. By contrast, in (c), each stage in our LAP-Net focuses on haze removal in different regions with different supervision, and after an adaptive weighted fusion, the final restoration exhibits a clear view on the whole. (Best viewed in color and zooming in.)

In the real-world situation, different positions of one image can be influenced by different haze conditions. Correspondingly, the degradation levels vary a lot along with different haze levels and the scene depth. As shown in Fig.1, regions far away from the camera like the buildings are degraded more than the close tree. That is because with the increasing of scene depth, the scattering mechanism becomes more complex. In the distant regions, the scene radiance should pass through more aerosol particles to the camera. It is likely to be scattered among particles for many times rather than reaching the camera after one time of scattering. Therefore, more efforts should be made to handle the different degradation between close and distant regions. To address this issue, the haze condition can be learned progressively with different stages of the network, and each stage only concerns one haze level. Specifically, mild haze can be tackled by the network with fewer stages, whereas heavy haze needs to be handled by more stages. Instead of giving the same supervision to each sub-network in the model, *different supervision* should be given to different stages, and in this way, the network can understand the gradual degradation caused by the haze in an easy-to-hard fashion. To this end, a sophisticated design of loss function that better guides the network to learn varying haze levels at multiple stages is required.

With the dehazed images at different stages, it becomes a critical problem on how to effectively integrate them to yield a natural result. Due to the complexity of real-world scenes, the restoration quality of different regions is not the same even in one stage. Thus we need to evaluate the restoration quality of different regions so that well-restored regions can contribute more to the final dehazed results. From Fig.1 we can see, if a region is clearly restored like the distant region in $\hat{\mathbf{J}}^S$ of our LAP-Net, the local variation of pixels at these positions will be higher. Such a variation can be estimated by local entropy. Therefore, we employ the measurement of local entropy as guidance. On the other hand, even if we are aware of whether a region in each stage is well-restored, we still need to know which stage can provide better results. It can be solved by an overall consideration of all stages. The prediction of haze level is also useful since it indicates which stage is likely to yield a better result in general. For example, if an image suffers from severe haze, the latter stages are more likely to produce a globally visually pleasant result. Therefore, we model the stage-wise relationship sequentially with the guidance of haze level prediction. Then the strengths of all the stages can be integrated to derive the final result.

Combining the above ideas, we propose an end-to-end model called level-aware progressive network (LAP-Net) for dehazing. With a specially designed loss function, each stage of LAP-Net focuses on one haze level and the entire work can learn the aggravation of haze with progressively increasing stages. After that, to effectively fuse the restoration in each stage, we design a hierarchical integration scheme. The lower level pays attention to the content of each stage with measuring the local entropy. It highlights the clearly restored regions for each stage. The higher level models the inter-stage relations from a global view. It weights all the stages with the guidance of haze level. With the integration scheme, we can preserve the good regional quality in each stage and fuse them for the final restoration.

Our contribution can be summarized as three-fold:

1) An end-to-end progressive dehazing network. Unlike the previous multi-stage methods [34, 35], stages in our network are supervised by the haze conditions from mild to dense orderly so that our network can dehaze progressively and is more adaptive to different haze conditions.

2) A hierarchical integration scheme guided by domain knowledge. The lower level of the scheme focuses the clear regions of each stage with the measurement of local entropy, while the higher level further updates the weight of each stage with the consideration of other stages and the guidance of haze level.

3) Extensive experiments that prove the integration of our designs can ultimately improve the performance of restoration both qualitatively and quantitatively.

2. Related work

Evolution of Dehazing approaches. Early methods are often based on the external information such as existing georeference model [19], user interaction parameter [32], and multiple images taken with different polarization degrees [37, 39] or weather conditions [31]. As it needs no extra devices or operations, single image dehazing techniques then raise the attention of researchers [41, 8, 14, 30, 2, 9, 54, 4, 11, 27, 36], among which the dark channel prior proposed by He *et al*. [14] is the most widely used. Recently, the rapid progress of deep learning also boosts learning-based dehazing methods. Tang *et al*. [42] first try to learn the transmission with a random forest model. The single-stage CNN [5] and multi-scale CNN [34] are proposed to estimate the transmission map. Then [20, 52] try to learn the two parameters of transmission and atmospheric light together. A conditional generative adversarial network (cGAN) [23] is also used to directly restore the haze-free image. Some methods [48, 28] dehaze by iteratively optimizing the transmission map.

Multi-stage strategy. Multi-stage networks have been widely used in both high-level and low-level tasks. Stacked hourglass and stage-wise refinement model are used for high-level issues like pose estimation [33, 49] or object detection [43]. For image de-weathering, Ren *et al*. [34, 35] employ a cascaded network to refine stage by stage. Yang *et al*. [50] repeatedly use a contextualized dilated network for removing rain streaks. Li *et al*. [24] combine the RNN and CNN structures to perform a stage-wise deraining. In most of these methods, the multi-stage network is used in a coarse-to-fine fashion. The latter stages always attempt to refine the features from its predecessor. In comparison, different stages in our network have *different responsibilities*, and they handle the haze in an easy-to-hard fashion. With a specifically designed loss, the early stages focus on mild haze, and the successive ones learn the denser haze based on the early ones. The results of them are integrated to exploit the advantages of each stage and thus to handle the complex real-world scenes.

Attention mechanism in reconstruction. As our guid-

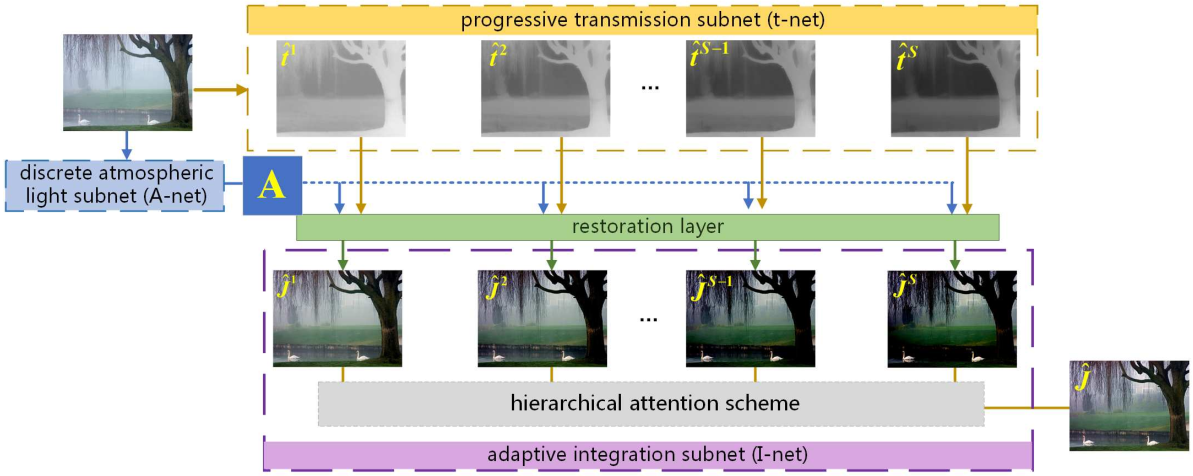

Figure 2. An overview of our method. We use the cascaded hourglass units to construct t-net, which is used for estimating the transmission map stage by stage with the supervision of different haze levels. Meanwhile, a residual dense pooling network serves as the A-net to learn the atmospheric light. The predictions of t-net and A-net are sent to the restoration layer to generate progressive restorations. Then we send the restored images of each stage as the input to the I-net for integration. The I-net cooperates with a hierarchical integration scheme that selects the clear regions of each stage and weights them with the guidance of haze level to restore the final haze-free image.

ance for integration pays attention to the local content and the sequential relation of different stages, we investigate the works using attention mechanism for reconstruction. The attention mechanism is not so common in low-level tasks as that in pose estimation [7], object detection [22] or tracking [6]. Zhang *et al*. [53] employ a channel-wise attention for super-resolution reconstruction. Xu *et al*. [46] propose an Attention-Gated Conditional Random Fields for detecting contours in images. In this study, considering the complexity of dehazing, our guidance is not learned with plain CNNs, but guided by the domain knowledge of haze level and local entropy. It can benefit both intra- and inter- stage restorations via a hierarchical integration scheme.

3. Proposed method

3.1. Atmospheric scattering model

The atmospheric scattering model is the basic model for most dehazing methods. It can be formulated as:

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

where x is the position of each pixel, $\mathbf{I}(x)$ and $\mathbf{J}(x)$ are the intensity of the hazy image and clear image, respectively. $\mathbf{A}$ is the atmospheric light, which indicates the overall environmental illuminance casting on the image. $t(x)$ denotes the medium transmission, which is the amount of scene radiance reaching the camera without being scattered by haze. In other words, $t(x)$ represents how much the haze influences on images. $t(x)$ is determined by the scene depth $d(x)$ and scattering coefficient β, and it can be expressed as:

$$t(x) = e^{-\beta d(x)}. \tag{2}$$

In general, the issue of dehazing first requires the estimation of $\mathbf{A}$ and $t(x)$. With these two parameters solved, the clear image $\mathbf{J}(x)$ can be obtained by reversing Eq.(1) as:

$$\mathbf{J}(x) = (\mathbf{I}(x) - \mathbf{A})/t(x) + \mathbf{A}. \tag{3}$$

Although the restoration can be theoretically realized with $t(x)$ and $\mathbf{A}$, it may not be practical for real-world scenarios. That is because in distant regions, the scattering may occur many times among the aerosol particles, and the degradation becomes more complex than what is expected in Eq.(2). In other words, β is not spatial-invariant across the image, and that hinders the accuracy of $t(x)$. Therefore, the accurate estimation of $t(x)$ is always a key step in existing works for improving the restoration quality. In our approach, $t(x)$ is estimated progressively by multi-stages according to the haze level, and the results are integrated via a hierarchical scheme for a natural restoration.

3.2. Overview of the proposed network

The overview of our level-aware progressive dehazing network is shown in Fig.2. It has four components: 1) the progressive transmission subnet (t-net), 2) the discrete atmospheric light subet (A-net), 3) the restoration layer and 4) the adaptive integration subnet (I-net). The hazy image is sent to t-net and A-net concurrently. Particularly, if the image is with color distortion, we first eliminate it according to [25]. Then the predictions $\{\hat{t}^s\}_{s=1}^S$ and $\hat{\mathbf{A}}$ pass the restoration layer to generate images with different restoration levels. The restorations are then weighted in I-net to yield the ultimate result. The whole network can be learned in an end-to-end manner.

In the **t-net**, the transmission $t(x)$ of the hazy image is learned by multiple progressive stages. Denote the number of stages by S and adopt $S = 4$ in experiments. Then we can obtain S stages of transmission maps. Instead of being supervised by the *same ground truth* like previous methods [34], different stages of the proposed network are supervised by *different maps*, which are generated with a fixed β value for each stage and represent an increasing haze level.

In the **A-net**, the atmospheric light is learned via a classification task. We discretize its interval and use the 18-layer residual network [16] for estimation.

The **restoration layer** obtains S dehazed images for the S progressive stages using Eq.(3). The inputs of this layer are transmission maps $t(x)$ from the t-net and the atmospheric light $\mathbf{A}$ from the A-net.

The **I-net** is proposed to combine the best-restored regions from the S stages and to derive the final result with the hierarchical integration scheme. The lower level considers the clearness of the content at each stage with local entropy. The higher level focuses on the information beyond the single stage. It takes sequential relations and the consistency of each stage with the haze level into consideration for updating weights.

3.3. The t-net for transmission map estimation

As aforementioned, previous methods [34, 5, 52] always learn the transmission with one ground truth corresponding to the hazy image. However, the uneven distribution of haze in real-world scenes makes the network hard to estimate the transmission in different haze conditions. Therefore, we propose a progressive t-net to estimate the transmission with different haze levels. This strategy helps the network handle the complex distribution of haze in real scenes. The t-net consists of S hourglass-shaped units for transmission estimation. Note that learning the transmission of mild haze is easier than that of dense haze. Therefore, we train the network in an easy-to-hard fashion, and the responsibility for estimating transmission is shared by a series of cascaded sub-networks. The prediction of transmission map $\hat{t}^s$ at stage s can be formulated as:

$$\hat{t}^s = \begin{cases} \mathcal{F}(\mathbf{I}, \theta^s) & s = 1 \\ \mathcal{F}(\mathbf{I}, \theta^s, \hat{t}^{s-1}) & s > 1, \end{cases} \tag{4}$$

where $\mathbf{I}$, $\hat{t}$ and s are the hazy image, the predicted transmission map and stage index, respectively. $\mathcal{F}$ denotes the network with the parameter θ^s at stage s. At the first stage, the t-net predicts the transmission map with a mild haze condition. Then the predicted transmission map and the hazy image are fed into the next stage to handle heavier haze.

3.4. The A-net for atmospheric light estimation

Atmospheric light is another important parameter in Eq.(1). In previous methods[14, 34, 5], the atmospheric light is manually obtained from the top 0.1% pixels in dark channel [14]. Such a statistical method cannot be directly integrated into the network. To make the haze removal process end-to-end learnable, we add a branch called A-net in the network for atmospheric light estimation.

Though the value of each component of $\mathbf{A}$ is continuous, we find that using MSE loss for learning has the regression-to-mean problem [44], which is unfavorable to obtain a precise estimation. Therefore, we consider the estimation of $\mathbf{A}$ as a classification task. Assuming the component of $\mathbf{A}$ to be a real value in $[A_l,\ A_h]$, we separate this interval into $n(A_h - A_l)$ discrete values (n is a precision controlling scalar and we have $n = 100$ here) and use the hazy image as the input for the classification. We employ the Res-18 network [16] as the basic network. The integration of low-level and high-level information can help estimate the global illuminance on images. Thus we add a multi-layer feature fusion module in the A-net. This module uses a global average pooling layer to normalize the features of different layers into resolution 1×1 and stack them together. Then it derives a normalized value to serve as the estimation of atmospheric light.

3.5. The I-net for integrating dehazed images from multiple stages

With the predicted $\{\hat{t}^s\}_{s=1}^S$ and $\hat{\mathbf{A}}$, a series of progressive dehazed results $\{\hat{\mathbf{J}}^s\}_{s=1}^S$ can be obtained via the restoration layer. Then we need to adaptively combine these results to obtain the ultimate restoration $\hat{\mathbf{J}}$, which can avoid the possible error propagation in the serial learning of $\hat{t}$ and further coordinate the good restoration of each stage. This process is formulated as:

$$\hat{\mathbf{J}} = \sum_s \mathbf{W}^s \circ \hat{\mathbf{J}}^s, \tag{5}$$

where $\mathbf{W}^s$ is the normalized three-channel weight map for stage s. It is extended by the one-channel weight w^s derived from the I-net. "$\circ$" denotes the element-wise multiplication. In this section, we discuss how we get the weights $\{w^s\}_{s=1}^S$ via the I-net.

3.5.1 The pipeline of I-net

The overall pipeline of the I-net is illustrated in Fig. 3. The derivation of w^s for each stage s involves three steps:

Step 1: Obtain the initial content-based weight w_c^s. w_c^s focuses on the content of $\hat{\mathbf{J}}^s$ at stage s. With the guidance of local entropy, we highlight the clearly restored regions of each stage. The confidence map of the clearness for each stage is used as the content-based prediction of w_c^s. The detail is discussed in Section 3.5.2.

Step 2: Obtain the intermediate contextual weight w_l^s from the initial content-based weights obtained in Step 1. w_l^s is derived according to the information beyond the content of stage s itself. We call that information as the contextual information since it includes the relationship with the previous/subsequent stages and the probability that stage s is consistent with the haze level. The deviation of w_l^s is detailed in Section 3.5.3.

Step 3: Obtain the final weight w^s by a pixel-wise refinement. w_l^s is refined with guided filter [15] and continuous CRF [47] to obtain the final weight w^s.

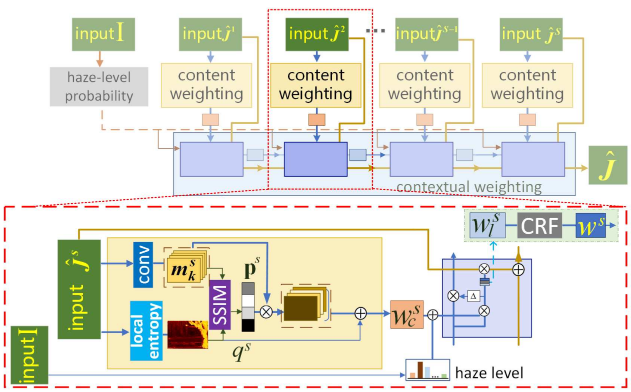

Figure 3. The structure of the I-net, which explores the memory network and a hierarchical scheme for adaptive integration.

3.5.2 Content-based weight prediction

The initial content-based weights w_c^s at stage s is obtained as following:

$$w_c^s = q^s \oplus (\sum_k p_k^s \circ m_k^s). \tag{6}$$

There are two kinds of inputs combined by element-wise addition $\oplus$ for w_c^s: one is *guidance weight* q^s obtained from the measurement of the local entropy; the other is *memory weight*, which consists of the sum of weighted candidate memories as $\sum_k p_k^s \circ m_k^s$. m_k^s is the memory storing the k-th feature map learned from $\hat{\mathbf{J}}^s$. It is weighted by a soft-addressing scheme associated with its probability p_k^s. $\circ$ here also denotes the element-wise multiplication.

Local entropy guidance weight q^s. The restoration of hazy image essentially needs a result with clear texture and vivid color. Local entropy can serve as the indicator of clearness since it evaluates the local variation of pixels. If the image is well-restored, the local variation in a patch can be dramatic. Otherwise, if the image is under-processed, the remaining haze causes the loss of clearness and leads to lower local variation. Similarly, the darkness caused by over-processing can also decrease the local variation. Therefore, we use the local entropy map as the guidance to focus on the clear patches, namely well-restored regions in each stage. Denote $\mathcal{N}$ as the set of neighboring pixels surrounding pixel location p in $\hat{\mathbf{J}}^s$, and $q^s(p)$ as the element at the pixel location p of the guidance weight q^s. $q^s(p)$ is calculated as the local entropy of $\mathcal{N}$ via a residual block as:

$$q^s(p) = \mathcal{G}(\mathrm{H}^s(\mathcal{N}), \theta_q^s) + \mathrm{H}^s(\mathcal{N}), \tag{7}$$

where $\mathcal{G}$ is a three-layer convolutional network with parameter θ_q^s. $\mathrm{H}^s(\mathcal{N}) = -\sum_{i\in\mathcal{N}} P_{\mathcal{N}}(x_i)\log P_{\mathcal{N}}(x_i)$ denotes the local entropy of $\mathcal{N}$ at stage s, which also filters the completely black or white pixels like [13], and x_i is one possible grayscale value in $\{x_1, x_2, \ldots, x_n\}$. In this paper we have $n = 256$. $P_{\mathcal{N}}(x_i)$ is the normalized histogram counts in the patch $\mathcal{N}$.

Memory m_k^s and its probability p_k^s. Besides q^s, we also hope to take the features from the restored image $\hat{\mathbf{J}}^s$ into consideration so that the predicted weight map can be smooth. To this end, we define K candidate memories of $\{m_k^s\}_{k=1}^K$ for each stage. Each memory is used to store the one-channel feature map extracted from $\hat{\mathbf{J}}^s$ via a network $\mathcal{H}$ as $m_k^s = \mathcal{H}(\hat{\mathbf{J}}^s, \theta_k^s)$. The probability p_k^s indicates the confidence that the feature map in m_k^s can serve as the weight. p_k^s here is defined as the matching degree between m_k^s and the local entropy map of H^s, which is calculated as follows:

$$p_k^s = \frac{\exp(\mathrm{SSIM}(m_k^s, \mathrm{H}^s))^\gamma}{\sum_k \exp(\mathrm{SSIM}(m_k^s, \mathrm{H}^s))^\gamma}, \tag{8}$$

where the SSIM$(\cdot)$ measures the matching degree and derived from the structural similarity index [45]. γ can amplify the focusing degree on one memory by enlarging it.

With m_k^s and p_k^s, we can obtain the memory weight. Like [12], instead of specifying single m_k^s with the highest probability as the final memory weight, we use the weighted sum, namely $\sum_k p_k^s \circ m_k^s$. This strategy not only makes the memory weight more comprehensive but more importantly makes this manipulation differentiable.

3.5.3 Contextual weight prediction

The initial content-based weight w_c^s focuses on regions inside one stage, but the relationship among the weights from different stages remains untreated. Take the restoration of $\hat{\mathbf{J}}^{S-1}$ and $\hat{\mathbf{J}}^S$ in Fig.2 as an example, if we only consider the content of the single stage, the buildings in distant regions are clearer than the others (like the trees in the close region) for both images. However, only when we consider from a contextual view, namely the relationship among the different stages, can we notice the distant region in stage S is restored better than that in stage $S-1$. Meanwhile, the guidance of haze level also plays an important role in balancing the stages. If the image is with dense haze, the latter restoration stages are more likely to recover clear details.

To obtain the comprehensive weights for each stage, we derive intermediate contextual weights from a higher level. We analyze the inter-stage relations and leverage the global guidance of haze level to highlight the restoration stage which has high consistency with the haze condition.

Inspired by [51], we employ a "budget gating" technique to model the intermediate contextual weights w_l^s based on the result of w_c^s. The contextual weighting is formulated as:

$$w_l^s = \begin{cases} w_c^s \oplus p_{lv}^s & s = 1 \\ \prod_{i=1}^{s-1}(1 - w_l^i)(w_c^s \oplus p_{lv}^s) & s > 1, \end{cases} \quad (9)$$

where $\prod$ denotes the element-wise product operator, which multiplies the inverse weight of all the early stages together with the weight at the current stage s. The inverse weight of w_l^i is obtained by $1 - w_l^i$. The p_{lv}^s is the consistency degree of stage s with the haze level of the input image, and it is obtained by learning with a classification module. This module has a similar architecture and learning strategy to A-net. To achieve the haze level classification, we first define S levels of the haze degree, which is in accord with the number of stages. Similar to the A-net, the hazy image is sent to the network but output S probabilities, which are normalized by the softmax function and serve as $\{p_{lv}^s\}_{s=1}^S$.

3.6. Network training

The entire network is trained in a two-stage scheme. In the first phase, the t-net and A-net are trained independently. The parameters $\{\theta^s\}_{s=1}^S$ of the t-net are learned by minimizing the MSE criterion $\mathcal{L}_t$ over N training samples:

$$\mathcal{L}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{S}||\mathcal{F}(\mathbf{I}_i, \theta^s, \hat{t}_i^{s-1}) - t_i^s||^2, \quad (10)$$

where $\mathcal{F}$ is the mapping function mentioned in Section 3.3, $\mathbf{I}_i$ is the i-th input hazy image from N samples and t_i^s is the s-th level transmission ground truth for the corresponding image. $\hat{t}_i^{s-1}$ is the prediction from the previous stage. Obviously, it only exists when s is over 2. The parameters for the A-net is learned via a normal cross-entropy loss $\mathcal{L}_A$:

$$\mathcal{L}_A = -\sum_{k=1}^{K} p_k log(\hat{p_k}), \quad (11)$$

where $p_k = \{p_1, p_2, \ldots, p_K\}$ is the ground-truth probability distribution of the k-th class of A value, and $\hat{p_k}$ is its estimation.

In the second phase, we initialize the t-net and A-net with parameters learned in the first phase, and jointly optimize the entire network with a multi-task loss function:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_A + \mathcal{L}_c, \quad (12)$$

where $\mathcal{L}_c$ is a restoration loss to minimize the difference between predicted haze-free image $\hat{\mathbf{J}}$ and ground truth $\mathbf{J}$ over N samples as $\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}||\hat{\mathbf{J}}_i - \mathbf{J}_i||^2$.

4. Experiments

4.1. Experiments setup

Network parameters. Our experiments are conducted with the Caffe framework [17] on a NVIDIA TITAN X GPU. During the training process, we randomly crop the input images with the size of 64×64. Then we send them into the network with a mini-batch size of 96. The CRF module parameters are the same as [47]. For the optimization, we adopt the ADAM algorithm [18]. The initial learning rate of the t-net is 0.001 and decreases by 10 times per 5000 iterations. The weight decay is 0.005, and β_1 and β_2 are fixed as the default value of 0.9 and 0.999. The training work stops after 30,000 iterations. The A-net is finetuned from the Resnet-18 model with an initial learning rate of 10^{-5}. The other settings are the same as those in the t-net. After the first training phase, we start to end-to-end train the whole network with the pre-trained parameters of the above two sub-networks. The learning rate of the layers that have been pre-trained is set to 10^{-7}. Except for them, the same hyper-parameters of the first phrase are all used in this phase.

Training data. Owing to the difficulty in obtaining realistic training data, we adopt the similar strategy as the previous methods [34, 5, 20, 35, 52] to synthesize hazy images with NYU depth dataset v2 [40] according to Eq.(2). Different from the previous ones, for better verifying the generalization of our method in real-world haze scenes, we do not select images from NYU dataset for testing. The scattering coefficient β is selected from $[0.4, 1.6]$, and the atmospheric light is from $[0.7, 1.0]$ to synthesize hazy samples.

4.2. Quantitative evaluation

We conduct the quantitative evaluation on two benchmark dataset of RESIDE [21] and O-HAZE [3]. The 500 outdoor image pairs of the Synthetic Objective Testing Set (SOTS) in RESIDE dataset are always with mild haze. The O-HAZE dataset contains 45 image pairs with denser haze than SOTS, and it can be used to test the performance for challenging conditions. Like [1], we use SSIM [45] and CIEDE2000 [38] for evaluation.

The first row in Table 1 reports the results of our method together with the other state-of-the-art ones on the outdoor image of SOTS in RESIDE dataset [1]. Ours performs favorably against the state-of-the-art dehazing methods. The metrics of SSIM and CIEDE2000 of our algorithm both achieve the best performance. These indexes are 0.019 and 0.336 better than the second-best one of [20] and [5], respectively. The visualized comparison in Fig.4 also shows ours can maintain the clear textures of close objects, like the bus and the ground while removing the influence of haze. The fidelity of color is also preserved and the restored image is not over-saturated like [14] or over-exposed like [52].

The comparison on the O-HAZE dataset is shown in the second row of Table 1. Facing the challenging condition of

[1]Note that there are two branches of indoor and outdoor in the latest released version of SOTS, which is a little different from what is stated in [21]. Our comparison is on the outdoor branch. For a fair comparison between the SOTS and O-HAZE, the metric of SSIM is calculated with the grayscale image, which is in accord with the results reported in [3].

Table 1. Quantitative comparison on RESIDE and O-HAZE dataset.

Dataset		DCP [14]	MSCNN [34]	DehazeNet [5]	AOD [20]	GFN [35]	DCPCN [52]	cGAN [23]	PDN [48]	Ours
RESIDE SOTS-outdoor	SSIM	0.786	0.855	0.873	0.915	0.836	0.883	0.881	0.865	**0.934**
	CIEDE2000	14.162	8.253	6.200	7.141	6.788	7.781	6.317	8.694	**5.864**
O-HAZE	SSIM	0.735	0.765	0.666	0.608	0.721	0.710	0.626	0.739	**0.798**
	CIEDE2000	20.745	14.670	17.348	19.110	17.269	22.311	14.162	17.402	**13.743**

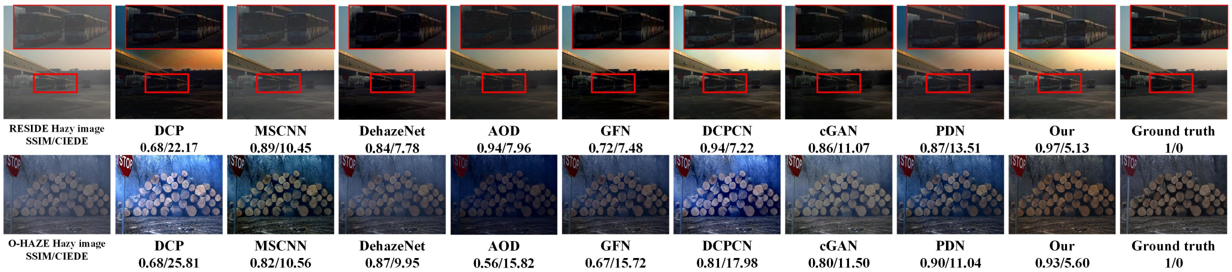

Figure 4. Visualization comparison of dehazing results on SOTS-outdoor of RESIDE [21] and O-HAZE [3]. SSIM/CIEDE2000 metrics are also marked below each image. (Best viewed in color and zooming in.)

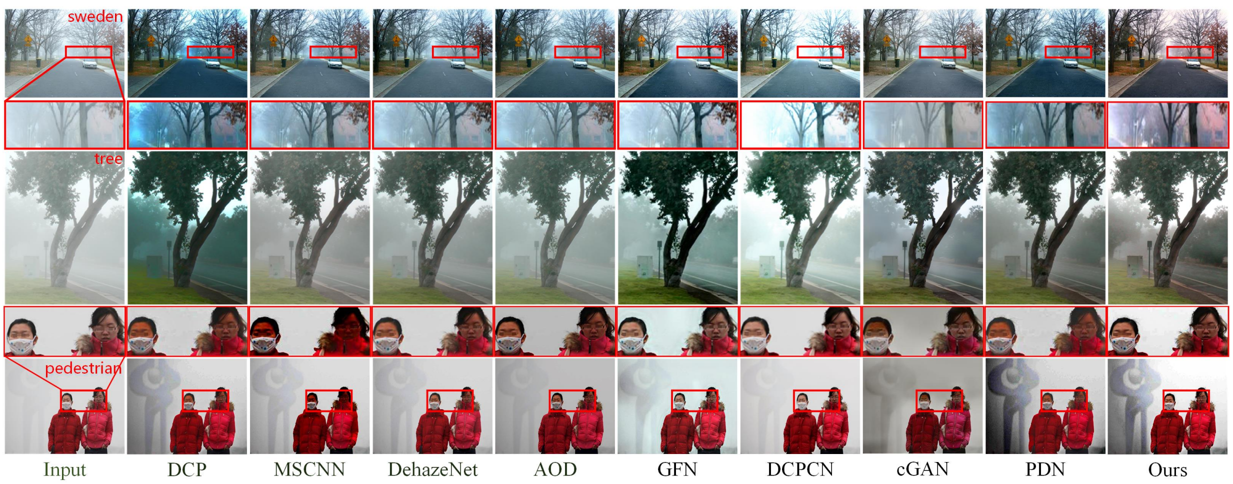

Figure 5. Qualitative comparison with other methods on real-world images. (Best viewed in color and zooming in.)

denser haze than SOTS, the performance of our network can still be better than existing works. Compared with MSCNN [35], which has the best performance of SSIM, the absolute improvement of our method is about 0.033. The CIEDE2000 is also 0.419 lower than the second-best method of cGAN [23], although it is trained on both indoor and outdoor data with perception loss and adversarial loss. From Fig.4 we can also find that our result preserves the global color fidelity of the haze-free image when compared with the others. The dense haze in the far distant regions is also restored without over-processing.

4.3. Visual Comparison

The qualitative comparison results of [14, 34, 5, 20, 35, 52, 23, 48] and ours are shown in Fig.5. With such a comparison, we find our method can achieve more appealing results under various conditions. The color and texture of distant objects, like the brick wall in "*sweden*", and the sculpture in "*pedestrian*" are restored more clearly in our result. Compared with methods like DCP [14] or DCPCN [52], our method also avoids over-saturation or over-exposure when restores the color, and it is attributed to our appropriate estimation of atmospheric light. Another remarkable strength of our method is that the good restoration of distant regions is not at the expense of the quality of close objects. Comparing the highlighted faces in "*pedestrian*", we find that ours keeps the color fidelity better than methods like DCP [14], MSCNN [34] and PDN [48], even though PDN restores the distant sculpture as well. It benefits from our adaptive integration scheme, which provides different restoration degrees to different regions according to the degradation condition.

4.4. Ablation study

In this section, we perform a study on the effect of each component of our method on O-HAZE dataset. Note that except for the compared part, the others are fixed as the final network.

Table 2. Quantitative comparison for ablation study.

single stage network		✓									
multi-stage network (coarse-to-fine)			✓								
progressive network	average fusion (baseline)			✓							
	LSTM				✓						
	CntWgt					✓			✓	✓	✓
	CtxtWgt-sequential balance						✓	✓	✓	✓	✓
	CtxtWgt-haze level							✓	✓	✓	✓
atmospheric light estimation	top 0.1%								✓		
	regression									✓	
	classification	✓	✓	✓	✓	✓	✓	✓			✓
SSIM		0.678	0.729	0.755	0.769	0.787	0.782	0.784	0.735	0.781	**0.798**
CIEDE2000		18.201	17.321	15.996	15.035	14.640	14.673	14.549	15.733	14.493	**13.743**

* CntWgt=content-based weighting, CtxtWgt=contextual weighting

Progressive strategy. As can be seen in Table 2, the single stage network gets the worst result. In comparison, the multi-stage network, even only trained in a coarse-to-fine fashion can improve the SSIM at 0.051 and CIEDE2000 at 0.88. It proves the multi-stage network performs better than the single-stage one. Compared with the coarse-to-fine network, the progressively trained network can always perform better no matter what the integrating strategy is. The basic one that averages all the stages improves metrics at 0.026 and 1.325 on the coarse-to-fine one, and it shows the effectiveness of our progressive learning strategy.

Integration scheme. The control group of LSTM scheme is employed to generate weights through a direct recurrent structure, and it can be deemed a kind of general attention mechanism-based strategy. We can see it improves the metrics at 0.014 and 0.961 on the baseline. When it turns to the domain knowledge-based attention with memory network, the improvement becomes more significant. Even only guided by local entropy, the metrics are improved at 0.032 and 1.356 on the baseline of averaging. For the contextual weighting, the global guidance of haze level can achieve slightly better performance than that of only considering the sequential relation. We also find that based on the result of content-based weighting, the contextual weighting can help to improve the performance. Otherwise, it is not as good as the content-based one. This phenomenon shows the correct prediction of each stage is the basis for the ultimate integration.

Atmospheric light estimation. The traditional one that estimates from the top 0.1% pixels in dark channel limits the performance since it is easily influenced by high-intensity objects. Compared with it, the A-subset trained with the MSE loss (regression) can substantially increase the performance at 0.046 and 1.24. Regarding the estimation as a classification task can make the estimation of $\mathbf{A}$ more precise and further improve the metrics at 0.017 and 0.75 on the "regression" one.

Effect of Losses. The effect of different loss terms in Eq.(11) is tested in Table 3. Since $\mathcal{L}_t$ and $\mathcal{L}_A$ are used to learn two parameters of the scattering model, it is not proper to test them separately. Therefore, we test the performance of $\mathcal{L}_t+\mathcal{L}_A$ and $\mathcal{L}_c$, respectively. With $\mathcal{L}_t+\mathcal{L}_A$, SSIM is 0.791 which is higher than the group with $\mathcal{L}_c$ at 0.014. It proves the scattering model can help to effectively restore the structure details of the hazy image. However, the CIEDE2000 is a little lower at 0.083 with $\mathcal{L}_c$, and it means $\mathcal{L}_c$ is more important to keep the color fidelity of the restoration.

Table 3. The effect of different terms of the loss function.

	SSIM	CIEDE2000
$\mathcal{L}_t+\mathcal{L}_A$ only	0.791	14.453
$\mathcal{L}_c$ only	0.777	14.370
$\mathcal{L}$	**0.798**	**13.743**

5. Conclusion

In this paper, we propose an end-to-end level-aware progressive network for single image dehazing. It first predicts the transmission and atmospheric light concurrently and outputs the restoration with progressively increasing dehazing levels. Then we deem the restorations as a sequence and employ a hierarchical scheme for computing the weight of each stage. To pick the clear regions of each stage and fuse them together, we propose a hierarchical integration scheme with domain knowledge of haze for addressing the stages in the network. Experimental results on two representative haze datasets and real-world images validate the effectiveness of our methods.

Acknowledgement The work was jointly supported by the National Key R&D Program of China under Grant No.2018YFC0807500, the National Key Research and Development Program of China No.238, the National Natural Science Foundations of China under grant No.61772396, 61472302, 61772392, the Fundamental Research Funds for the Central Universities under grant No. JB170306, JB170304, JBF180301, and Xi'an Key Laboratory of Big Data and Intelligent Vision under grant No.201805053ZD4CG37.

References

[1] Cosmin Ancuti, Codruta Orniana Ancuti, and Christophe De Vleeschouwer. D-hazy: A dataset to evaluate quantitatively dehazing algorithms. In *ICIP*, pages 2226–2230. IEEE, 2016.

[2] Codruta Orniana Ancuti and Cosmin Ancuti. Single image dehazing by multi-scale fusion. *TIP*, 22(8):3271–3282, 2013.

[3] Codruta Orniana Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *CVPRW*, 2018.

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

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

[6] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *ICCV*, pages 4836–4845, 2017.

[7] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, pages 1831–1840, 2017.

[8] Raanan Fattal. Single image dehazing. In *SIGGRAPH*, pages 1–9, 2008.

[9] Raanan Fattal. Dehazing using color-lines. *ACM TOG*, 34(1):1–13, 2014.

[10] Chen Feng, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Sabine Süsstrunk. Near-infrared guided color image dehazing. In *ICIP*, pages 2363–2367, 2013.

[11] Adrian Galdran, Aitor Alvarez-Gila, Alessandro Bria, Javier Vazquez-Corral, and Marcelo Bertalmło. On the duality between retinex and image dehazing. In *CVPR*, 2018.

[12] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[13] Nicolas Hautiere, Jean-Philippe Tarel, Didier Aubert, and Eric Dumont. Blind contrast enhancement assessment by gradient ratioing at visible edges. *Image Anal. Stereol.*, 27(2):87–95, 2011.

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *CVPR*, pages 1956–1963. IEEE, 2009.

[15] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *ECCV*, pages 1–14. Springer, 2010.

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

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678. ACM, 2014.

[18] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] 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 TOG*, 27(5), 2008.

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

[21] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2019.

[22] Hongyang Li, Yu Liu, Wanli Ouyang, and Xiaogang Wang. Zoom out-and-in network with map attention decision for region proposal and object detection. *IJCV*, 127(3):225–238, 2019.

[23] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. Single image dehazing via conditional generative adversarial network. In *CVPR*, 2018.

[24] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 262–277. Springer, 2018.

[25] Yunan Li, Qiguang Miao, Jianfeng Song, Yining Quan, and Weisheng Li. Single image haze removal based on haze physical characteristics and adaptive sky region detection. *Neurocomputing*, 182(3):221–234, 2016.

[26] Yu Li, Shaodi You, Michael S Brown, and Robby T Tan. Haze visibility enhancement: A survey and quantitative benchmarking. *CVIU*, 165:1–16, 2017.

[27] Zhengguo Li and Jinghong Zheng. Single image de-hazing using globally guided image filtering. *TIP*, 27(1):442–450, 2018.

[28] Qi Liu, Xinbo Gao, Lihuo He, and Wen Lu. Single image dehazing with depth-aware non-local total variation regularization. *TIP*, 27(10):5178–5191, 2018.

[29] Earl J McCartney. *Optics of the atmosphere: scattering by molecules and particles*. John Wiley and Sons, Inc., New York, 1976.

[30] Gaofeng Meng, Ying Wang, Jiangyong Duan, Shiming Xiang, and Chunhong Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *ICCV*, pages 617–624. IEEE, 2013.

[31] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *IJCV*, 48(3):233–254, 2002.

[32] Srinivasa G Narasimhan and Shree K Nayar. Interactive (de) weathering of an image using physical models. In *ICCVW*, 2003.

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

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

[35] 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*, 2018.

[36] Sanchayan Santra, Ranjan Mondal, and Bhabatosh Chanda. Learning a patch quality comparator for single image dehazing. *TIP*, 27(9):4598–4607, 2018.

[37] Yoav Y Schechner, Srinivasa G Narasimhan, and Shree K Nayar. Polarization-based vision through haze. *Applied Optics*, 42(3):511–525, 2003.

[38] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[39] Sarit Shwartz, Einav Namer, and Yoav Y Schechner. Blind haze separation. In *CVPR*, pages 1984–1991. IEEE, 2006.

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[41] Robby T Tan. Visibility in bad weather from a single image. In *CVPR*, pages 1–8, 2008.

[42] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *CVPR*, pages 2995–3002. IEEE, 2014.

[43] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017.

[44] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018.

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[46] Dan Xu, Wanli Ouyang, Xavier Alameda-Pineda, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In *NeurIPS*, pages 3961–3970, 2017.

[47] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017.

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

[49] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017.

[50] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017.

[51] Chris Ying and Katerina Fragkiadaki. Depth-adaptive computational policies for efficient visual tracking. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 109–122. Springer, 2017.

[52] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018.

[53] 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.

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