# DesnowNet: Context-Aware Deep Network for Snow Removal

Yun-Fu Liu, *Member, IEEE*, Da-Wei Jaw, Shih-Chia Huang, *Senior Member, IEEE*, and Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*— Existing learning-based atmospheric particle-removal approaches such as those used for rainy and hazy images are designed with strong assumptions regarding spatial frequency, trajectory, and translucency. However, the removal of snow particles is more complicated because they possess additional attributes of particle size and shape, and these attributes may vary within a single image. Currently, hand-crafted features are still the mainstream for snow removal, making significant generalization difficult to achieve. In response, we have designed a multistage network named DesnowNet to in turn deal with the removal of translucent and opaque snow particles. We also differentiate snow attributes of translucency and chromatic aberration for accurate estimation. Moreover, our approach individually estimates residual complements of the snow-free images to recover details obscured by opaque snow. Additionally, a multi-scale design is utilized throughout the entire network to model the diversity of snow. As demonstrated in the qualitative and quantitative experiments, our approach outperforms state-of-the-art learning-based atmospheric phenomena removal methods and one semantic segmentation baseline on the proposed Snow100K dataset. The results indicate our network would benefit applications involving computer vision and graphics.**

***Index Terms*— Snow removal, deep learning, convolutional neural networks, image enhancement, image restoration.**

## I. Introduction

ATMOSPHERIC phenomena, e.g., rainstorms, snowfall, haze, or drizzle, obstructs the recognition of computer vision applications. Such conditions can influence sensitive usages such as intelligent surveillance systems and result in higher risks of false alarms and unstable machine interpretation. Figure 1 shows a subjective but quantifiable result for a concrete demonstration, in which the labels and corresponding confidences are both supported by Google Vision API.[1]

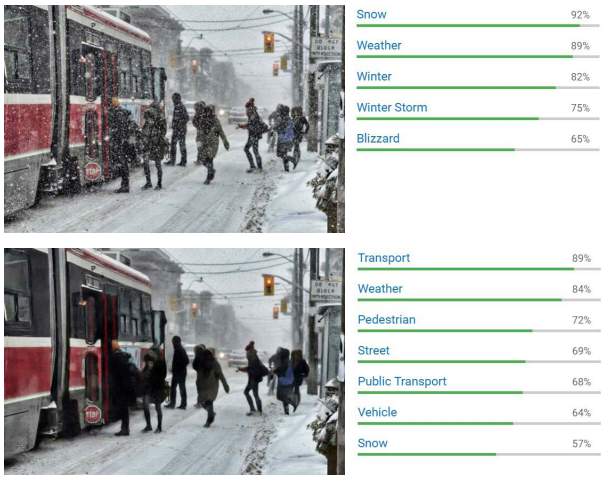



Fig. 1. Realistic winter photograph (top) and corresponding snow removal result (bottom) of the proposed method with labels and confidences supported by Google Vision API.

It illustrates that these atmospheric particles could impede the interpretation of object-centric labeling - in this case, of *pedestrian* and *vehicle*.

Numerous atmospheric particle removal techniques have been proposed to eliminate object obscuration by particles and thus provide richer details. To this end, removing haze and rain particles from images are currently common topics of research. So far, approaches for dealing with haze particles have been designed with a strong assumption [1]–[3], attenuation prior [4]–[10] and feature learning [11], [12], based on the observation that haze is uniformly accumulated over an entire image. Moreover, rain removal techniques aim to model general characteristics such as edge orientations [13]–[16], shapes [17] or patterns [18] to detect and remove rain particles. In recent years, learning-based approaches [19]–[21] have attracted much attention as they are purportedly more effective than previous hand-crafted features due to their significantly improved generalization abilities.

Although the learning-based rain and haze removal methods are capable of locating and removing atmospheric particles, it is hard to adapt them for snow removal because of snow's complex characteristics. Specifically, its uneven density, diversified particle sizes and shapes, irregular trajectory and transparency make snow removal a more difficult task to accomplish and inapplicable for other learning-based methods. Meanwhile, existing snow removal approaches focus

Manuscript received August 16, 2017; revised January 5, 2018; accepted January 26, 2018. Date of publication February 14, 2018; date of current version March 29, 2018. This work was supported by the Ministry of Science and Technology of the Republic of China under Grant MOST 106-2221-E-027-017-MY3, Grant MOST 106-2221-E-027-126-MY2, and Grant MOST 105-2923-E-027-001-MY3. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Shih-Chia Huang.)*

Y.-F. Liu is with Alibaba DAMO Academy, Hangzhou 311121, China (e-mail: yunfuliu@gmail.com).

D.-W. Jaw and S.-C. Huang are with the Department of Electronic Engineering, National Taipei University of Technology, Taipei 106, Taiwan (e-mail: jdw.davidjaw@gmail.com; schuang@ntut.edu.tw).

J.-N. Hwang is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: hwang@uw.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2806202

[1]Google Vision API: https://cloud.google.com/vision/

on designing hand-crafted features, which often results in a weak generalization ability such as that which has occurred in the fields of rain and haze removal.

In this paper, we design a network named *DesnowNet* to in turn deal with complicated translucent and opaque snow particles. The reason behind this multistage design is to acquire more recovered image content by which to accurately estimate and restore details lost to opaque snow particle coverage. In addition, we model snow particles with the attributes of a snow mask, which considers only the translucency of snow at each coordinate, as well as an independent chromatic aberration map by which to depict subtle color distortions at each coordinate for better image restoration. Because of the variations in size and shape of snow particles, the proposed approach comprehensively interprets snow through context-aware features and loss functions. Experimental results demonstrate that the proposed approach yields a significant improvement of prediction accuracy as evaluated via the proposed Snow100K dataset.

The rest of this paper is organized as follows: Section II provides an overview of the existing atmospheric particle removal methods. Sections III and IV elaborate the details of the proposed DesnowNet and Snow100K dataset, respectively. Finally, Section V presents the experimental results, and Section VI draws the conclusions.

## II. Related Works

### A. Rain Removal

Kang *et al*. [13] proposed the first image decomposition framework for single image rain streak removal. Their method is based on the assumption that rain streaks have similar gradient orientations in one image, as well as high spatial frequency. These features are considered to segment the rainy component with sparse representation for the removal process. Luo *et al*. [16] proposed a similar discriminative sparse coding methodology with the same assumption to remove rain particles from rainy images. However, both methods introduce stripe-like artifacts.

To cope with this, Son and Zhang [17] designed a shrinkage-based sparse coding technique to improve visual quality. Chen *et al*. [15] extended [13] by employing the histogram of oriented gradients feature (HOG [22]), depth of field, and Eigen color to better separate rain components from others. Li *et al*. [18] proposed an image decomposition framework using the Gaussian mixture model to accomplish rain removal by accommodating the similar patterns and orientations among rain particles.

To further improve generalization, Fu *et al*. [20] and [21] and Yang *et al*. [19] proposed convolutional neural network (CNN)-based methods to remove rain particles. In the design of [20] and [21], they first separate a rainy image into a high frequency detail layer and a low frequency base layer, under the assumption that they possess and are devoid of rain particles, respectively. The design of Yang *et al*.'s JORDER method [19] utilizes a contextual dilated network to extract features with receptive fields of three different sizes. It also yields a residual complement from predicted priors and extracted features to enhance the resulting visibility. Those method outperforms the former hand-crafted methods when it comes to removal accuracy and the clarity of the results. Although individual cases of snowy images may be similar in appearance to those featuring rain streaks, the most common scenario in which a snowy image features coarse-grained snow particles will result in failure due to the overly focused spatial frequency of [20] and [21], and the lack of ability to recover the completely opaque particles of [19].

### B. Haze Removal

Tan's method [1] assumes that haze-free images possess better contrast than images with haze. Thus, they remove haze particles by maximizing the local contrast of hazy images. Fattal [2] removed haze by estimating the albedo of a scene and inferring the transmission medium in hazy images. Chen and Huang [4] proposed a self-adjusting transmission map estimation technique by taking advantage of edge collapse. He *et al*. [5] accomplish haze removal via the dark channel prior, attracting considerable attention from their effective results.

Chen *et al*. [6] proposed a gain intervention refinement filter to speed up the runtime of [5]. In [7], a hybrid dark channel prior is proposed to avoid the artifacts introduced by localized light sources. As an extension, the dual dark channel prior with adaptable localized light detection [8] was introduced to automatically adjust the size of a binary mask to conceal localized light sources and achieve more reliable results for haze removal. Huang *et al*. [9] proposed a Laplacian-based visibility restoration technique to refine the transmission map and solve color cast issues. In [10], a transmission map refinement procedure is proposed to avoid complex structure halo effects by preserving the edge information of the image.

However, all these hand-crafted methods frequently suffer from similar failure cases due to haze-relevant priors or heuristic cues. Most recently, learning-based method [11] and [12] focused on learning the mapping between a hazy image and its corresponding medium transmission map. In their atmospheric scattering model, they assume that a clear image is a superposition of an equal-brightness haze mask with different translucency and a clear image. While their method is effective in some cases, there are limitations. For instance, the brightness of the haze mask is directly determined by the global maximum of the image, which is not learnable and may also cause problems for generalization. Moreover, their approach also assumes that haze particles are translucent and that the images lack opaque corruption. Such architecture focusing on extracting global features is designed to recover entire image corruption, which leads to inapplicable on snow removal.

### C. Snow Removal

Unlike the characteristics of atmospheric particles in rainy and hazy images that might be described as relatively similar in spatial frequency, trajectory, and translucency, the variations in the particle shape and size of snow make it more complicated. However, existing snow removal methods inherited

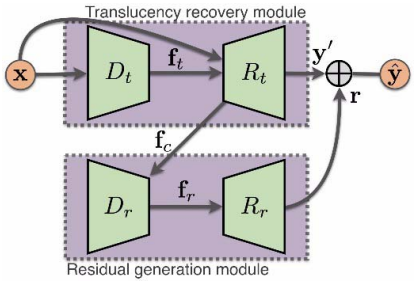



Fig. 2. Overview of the proposed DesnowNet, where the operator $\oplus$ denotes the element-wise addition as depicted in Eq. (2).

priors of rainfall-driven features (e.g., HOG) [23], [24], frequency space separation [14], [25] and color assumptions [26], [27] to model falling snow particles. These features not only model just the partial characteristics of snow, but worsen the prospects of generalization.

## III. Proposed Method

We now describe our proposed method for removing falling snow particles from snowy images. Suppose that a snowy color image $\mathbf{x} \in [0, 1]^{p\times q\times 3}$ of size $p \times q$ is composed of a snow-free image $\mathbf{y} \in [0, 1]^{p\times q\times 3}$ and an independent snow mask $\mathbf{z} \in [0, 1]^{p\times q\times 1}$ which introduces only the translucency of snow. This relationship can be described as:

$$\mathbf{x} = \mathbf{a} \odot \mathbf{z} + \mathbf{y} \odot (1 - \mathbf{z}) \tag{1}$$

where $\odot$ denotes element-wise multiplication, and the chromatic aberration map $\mathbf{a} \in \mathbb{R}^{p\times q\times 3}$ introduces the color aberration at each coordinate.

To derive the estimated snow-free image $\hat{\mathbf{y}}$ from a given $\mathbf{x}$, the estimation of an accurate snow mask $\hat{\mathbf{z}}$ as well as a chromatic aberration map $\mathbf{a}$ are crucial to achieve an appealing visual quality. To this end, we design a network with multi-scale receptive fields to model all of the variations of snow particles as introduced in Section II-C. Specifically, the proposed network as depicted in Fig. 2 consists of two modules for different purposes: 1) the translucency recovery (TR) module, which recovers areas obscured by translucent snow particles; 2) the residual generation (RG) module, which generates the residual complement $\mathbf{r} \in \mathbb{R}^{p\times q\times 3}$ of the estimated snow-free image $\mathbf{y}'$ for portions completely obscured by opaque snow particles according to the clear (non-covered) areas as well as those recovered by the TR module. Notably, unlike utilizing the hidden features in the TR module to yield residual complement as in Yang *et al.*'s JORDER method [19], our RG module focuses on interpreting the residual from the snow-free image $\mathbf{y}'$ and the estimated priors $\mathbf{a}$ and $\mathbf{z}$.

The RG module not only refines the result from the TR module, but also can benefit the optimization of the two-stage network as depicted in [28]. To do so, two modulized descriptors $D_t$ and $D_r$ are utilized to extract multi-scale features $\mathbf{f}_t$ and $\mathbf{f}_r$ with varied receptive fields, respectively, and two modulized recoverers $R_t$ and $R_r$ are designed to yield the snow-free estimate $\mathbf{y}'$ and residual complement $\mathbf{r}$, respectively. Thus, the estimated snow-free result ($\hat{\mathbf{y}}$) can be derived as described below:

$$\hat{\mathbf{y}} = \mathbf{y}' + \mathbf{r} \tag{2}$$

The derivation of $\mathbf{y}'$ and $\mathbf{r}$ are further described in the following subsections. Notably, due to the possibility that the value range of $\mathbf{r}$ could make $\hat{\mathbf{y}}$ exceed the expected range $[0, 1]$, we clip when $\hat{y}_i > 1$ and $\hat{y}_i < 0$ during inference yet keep it unchanged during training, where $\hat{y}_i \in \hat{\mathbf{y}}$. Fig. 3 illustrates examples of the variables in Fig. 2 to aid comprehension of this process.

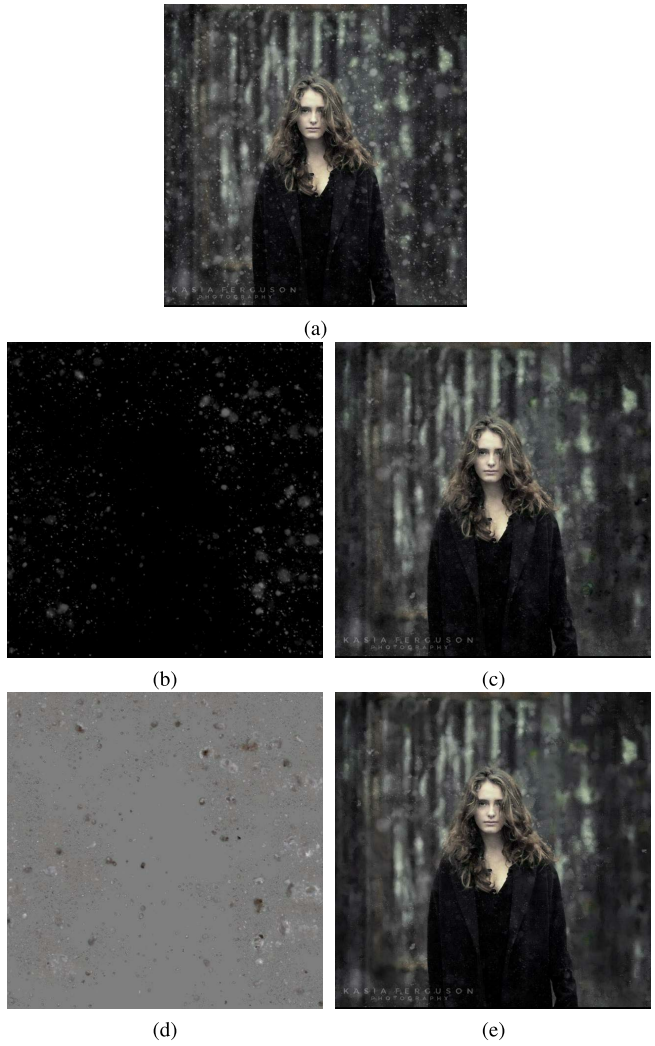



Fig. 3. Results of DesnowNet during inference, where the gray color in (d) denotes $r = 0$, and white and black colors represent positive and negative values, respectively; $r \in \mathbf{r}$. (a) Snowy image ($\mathbf{x}$). (b) Aberrated snow mask ($\mathbf{a} \odot \hat{\mathbf{z}}$). (c) Estimated snow-free output ($\mathbf{y}'$). (d) Residual complement ($\mathbf{r}$). (e) Estimated snow-free output ($\hat{\mathbf{y}}$).

### A. Descriptor

Both $D_t$ and $D_r$ (illustrated in Fig. 2) utilize the same type of descriptor to introduce the variations of snow particles. Specifically, although the purposes of $D_t$ and $D_r$ are inherently different as described above, they share the same need for resolving the variations in snow particles. To this end, we first employ Inception-v4 [29] as the backbone due

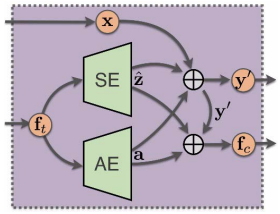



Fig. 4. Overview of the $R_t$ submodule, where the upper operator $\oplus$ is defined in Eq. (5) and the lower operator $\oplus$ denotes the concatenate operation.

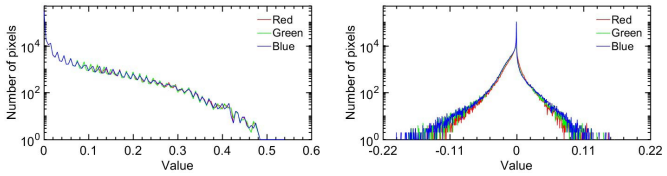



Fig. 5. Histograms of the results of Fig. 3(b) and (d), where RGB represents the color channels. Best viewed in color.

to its highly optimized features at multi-scale receptive fields, a network whose success has been demonstrated in diverse studies [30], [31].

To further extend context-awareness and exploit multi-scale features, as inspired by the atrous spatial pyramid pooling (ASPP) in DeepLab [32], we proposed a subnetwork termed dilation pyramid (DP) as defined below,

$$\mathbf{f}_t = \Big\|_{n=0}^{\gamma} B_{2^n}(\Phi(\mathbf{x})) \tag{3}$$

where $\Phi(\mathbf{x})$ represents the features output from the last convolution layer of Inception-v4 [29] with a given image $\mathbf{x}$, $B_{2^n}(\cdot)$ denotes the dilated convolution [33] (the atrous convolution in [32]) with dilation factor $2^n$, $||$ represents the concatenating operation, $\gamma \in \mathbb{R}$ denotes the levels of the dilation pyramid, and $\mathbf{f}_t$ (or $\mathbf{f}_r$ for $D_r$ in Fig. 2) denotes the output feature of $D_t$. By utilizing multiple dilated convolutions, DP can further enhance the capability of extracting scaling invariant features, which is crucial for falling snow. However, although both of JORDER [19] and the ASPP in DeepLab [32] utilize the same dilated convolution, we concatenate the extracted features instead of their summing up operations to preserve the property at each scale.

### *B. Recovery Submodule*

**Pyramid maxout.** The recovery submodule ($R_t$) of the translucency recovery (TR) module as illustrated in Fig. 2 generates the estimated snow-free output ($\mathbf{y}'$) by recovering the details behind translucent snow particles. To accurately model this variable, we physically and separately define it as two individual attributes as depicted in Fig. 4: 1) snow mask estimation (SE), which generates a snow mask ($\hat{\mathbf{z}} \in [0,1]^{p\times q\times 1}$) to model the translucency of snow with a single channel, and 2) aberration estimation (AE), which generates the chromatic aberration map ($\mathbf{a} \in \mathbb{R}^{p\times q\times 3}$) to estimate the color aberration of each RGB channel. Their relationship is defined in Eq. (1). Due to the variation in size and shape of snow particles, we design an emulation architecture termed *pyramid maxout* as defined below to enforce the network to select a robust feature map from different receptive fields as its output. It is similar to the general form of maxout [34] in terms of its competitive policy, and it is defined as:

$$M_\beta(\mathbf{f}_t) = max(\text{conv}_1(\mathbf{f}_t), \text{conv}_3(\mathbf{f}_t), \ldots, \text{conv}_{2\beta-1}(\mathbf{f}_t)) \tag{4}$$

where $\text{conv}_n(\cdot)$ denotes the convolution with kernel of size $n \times n$, $max(\cdot)$ denotes the element-wise maximum operation, $\mathbf{f}_t$ has been defined in Eq. (3), and parameter $\beta \in \mathbb{R}$ controls the scale of the pyramid operation. In our design, SE and AE utilize the same architecture as defined in Eq. (4) to generate $\hat{\mathbf{z}}$ and $\mathbf{a}$, respectively.

**Translucency recovery (TR).** $R_t$ in TR module recovers the content behind translucent snow to yield an estimated snow-free image $\mathbf{y}'$. The relationship can be formulated with Eq. (1):

$$y'_i = \begin{cases} \frac{x_i - a_i \times \hat{z}_i}{1-\hat{z}_i}, & \text{if } \hat{z}_i < 1 \\ x_i, & \text{if } \hat{z}_i = 1 \end{cases} \tag{5}$$

where the subscript $i$ denotes the $i$-th element in corresponding matrices. Notably, a condition of $y'_i = x_i$ occurs if $\hat{z}_i = 1$ addresses the special case of nonexistent $y'_i$ as defined in Eq. (1). Fig. 3(b) and 5(a) show the results of $\mathbf{a} \odot \hat{\mathbf{z}}$ as defined in Eq. (5). Although the snow particles are normally presented in middle gray in most photographs, such as that shown in Fig. 3(b), our chromatic aberration map $\mathbf{a}$ describes potential subtle color variations as depicted in Fig. 5(a). This not only further complements the property of colored snow particles, but improves generalization of variation in ambient light.

**Residual generation (RG).** $R_r$ serves a different purpose that generates a residual complement $\mathbf{r} \in \mathbb{R}^{p\times q\times 3}$ for the estimated snow-free image $\mathbf{y}' \in \mathbb{R}^{p\times q\times 3}$ to yield a visually appealing $\hat{\mathbf{y}}$ as formated in Eq. (2). To this end, the RG module needs to know the explicit locations of estimated snow particles in $\hat{\mathbf{z}}$, the corresponding chromatic aberration map $\mathbf{a}$, and the recovered output $\mathbf{y}'$. The residual complement can be formulated as:

$$\begin{aligned} \mathbf{r} &= R_r(D_r(\mathbf{f}_c)) \\ &= \Sigma_\beta(\mathbf{f_r}) = \sum_{n=1}^{\beta} \text{conv}_{2n-1}(\mathbf{f_r}) \end{aligned} \tag{6}$$

where we set $\mathbf{f}_c = \mathbf{y}' || \hat{\mathbf{z}} || \mathbf{a}$ in order to allow the RG module to explore the potential relevance among these semantic features, $\mathbf{f}_r$ denotes the output of $D_r$ as depicted in Fig. 2, $\beta$ is identical to that of in Eq. (4), and the pyramid sum $\Sigma_\beta(\cdot)$ aggregates the multi-scale features to model the variation in snow particles.

It is worth noting that, we do not consider the pyramid maxout as defined in Eq. (4) useful for predicting $\mathbf{r}$ for two reasons: 1) the distribution of $\mathbf{r}$ should be a zero-mean distribution as shown in Fig. 5(b) to fairly compensate for the estimated snow-free result $\mathbf{y}'$, and the pyramid maxout would violate this property; 2) pyramid maxout focuses on

the robustness of the estimated features over various scales, whereas the pyramid sum aims to simulate visual perception at all scales simultaneously.

### C. Loss Function

Although the primary purpose of this work aims to remove falling snows from snowy images ($\mathbf{x}$) and thereby approach the snow-free ground truth ($\mathbf{y}$), perceptual loss is needed to simulate its visual similarity. Element-wise Euclidean loss is often considered for this purpose. However, it simply evaluates feature representation at single scale constraints for the simulation of various viewing distances pertinent to human vision, and unfortunately introduces unnatural artifacts. To address this issue, Johnson *et al*. [35] constructed a loss network that measures losses at certain layers. We adopt this same idea, retaining its contextual features while formulating it as a lightweight pyramid loss function as defined below:

$$\mathfrak{L}(\mathbf{m}, \hat{\mathbf{m}}) = \sum_{i=0}^{\tau} \| P_{2^i}(\mathbf{m}) - P_{2^i}(\hat{\mathbf{m}}) \|_2^2 \tag{7}$$

where $\mathbf{m}$ and $\hat{\mathbf{m}}$ denote two images of the same size, $\tau \in \mathbb{R}$ denotes the level of loss pyramid, $P_n$ denotes the max-pooling operation with kernel size $n \times n$ and stride $n \times n$ for non-overlapped pooling.

Two feature maps $\mathbf{y}'$ and $\hat{\mathbf{y}}$ represent the estimated snow-free images as illustrated in Fig. 2. In addition, the individual estimated snow mask $\hat{\mathbf{z}}$ depicted in Fig. 4 represents the perceptual translucency of snows in the ideal snow-free image $\mathbf{y}$. Hence, the overall loss function is defined as:

$$\mathcal{L}_{overall} = \mathcal{L}_{\mathbf{y}'} + \mathcal{L}_{\hat{\mathbf{y}}} + \lambda_{\hat{\mathbf{z}}} \mathcal{L}_{\hat{\mathbf{z}}} + \lambda_{\mathbf{w}} \|\mathbf{w}\|_2^2 \tag{8}$$

where $\lambda_{\hat{\mathbf{z}}} \in \mathbb{R}$ denotes the weighting to leverage the importance of snow mask $\hat{\mathbf{z}}$ and both snow-free estimates $\mathbf{y}'$ and $\hat{\mathbf{y}}$, where $\mathbf{y}'$ and $\hat{\mathbf{y}}$ are set at equal importance; $\lambda_{\mathbf{w}} \in \mathbb{R}$ denotes the weighting to the $l2$-norm regularization, $\mathbf{w}$ denotes the weighting of the entire DesnowNet; the losses of $\mathbf{y}'$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are defined as:

$$\mathcal{L}_{\hat{\mathbf{z}}} = \mathfrak{L}(\mathbf{z}, \hat{\mathbf{z}}), \mathcal{L}_{\hat{\mathbf{y}}} = \mathfrak{L}(\mathbf{y}, \hat{\mathbf{y}}), \text{and } \mathcal{L}_{\mathbf{y}'} = \mathfrak{L}(\mathbf{y}, \mathbf{y}'), \tag{9}$$

where $\mathbf{z}$ and $\mathbf{y}$ denote the ground truth of the snow mask and the ideal snow-free image, and $\mathfrak{L}(\cdot)$ was defined in Eq. (7).

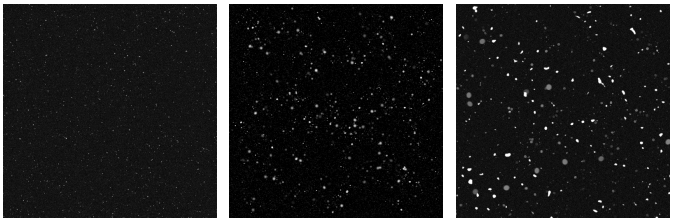
## IV. Dataset

The Snow100K dataset[2] $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}^N$ consists of 1) 100k synthesized snowy images ($\mathbf{x}_i$), 2) corresponding snow-free ground truth images ($\mathbf{y}_i$), 3) snow masks ($\mathbf{z}_i$) and 4) 1,329 realistic snowy images. The images of 2) and 4) were downloaded via the Flickr api, and were manually divided into snow and snow-free categories, respectively. In addition, we normalized the size of the largest boundary of each image to 640 pixels and retained its original aspect ratio.

To synthesize the snowy images, we produced 5.8k base masks via PhotoShop to simulate the variation in the particle sizes of falling snow. Each base mask consists of one of the small, medium, and large particle sizes. Meanwhile, the snow particles in each base mask feature different densities, shapes, movement trajectories, and transparencies, by which to increase the variation. The number of base masks in each category is exhibited in Table I. Corresponding examples are shown in Fig. 6.

TABLE I
NUMBER OF BASE MASKS IN Snow100K DATASET

| Particle size | Small | Medium | Large |
|---|---|---|---|
| # masks | 3,800 | 1,600 | 400 |



(a) (b) (c)

Fig. 6. Example of base masks with differing particle sizes. (a) Small. (b) Medium. (c) Large.

TABLE II
NUMBER OF SAMPLES IN EACH SUBSET OF Snow100K DATASET

| Subset | Snow100K-S | Snow100K-M | Snow100K-L |
|---|---|---|---|
| Training | 16,643 | 16,622 | 16,735 |
| Test | 16,611 | 16,588 | 16,801 |

Three subsets of synthesized snowy images $\mathbf{x}_i$ are organized for different snowfalls. Notably, the base masks in Table I are utilized here for simulation: 1) Snow100K-S: samples are only overlapped with one randomly selected base mask from the *Small* category 2) Snow100K-M: samples are overlapped with one randomly selected base mask from the *Small* category and one base mask from the *Medium* category; 3) Snow100K-L: three base masks randomly selected from each of the three categories are adopted to synthesize samples in this subset. The numbers of synthesized snowy samples in each category are exhibited in Table II. For the superposition process, we append two additional random factors to further increase the randomness for generalization. These include 1) Snow brightness: The brightness of the superposed snow in base mask is randomly set within $[max(\mathbf{y}_i) \times 0.7, max(\mathbf{y}_i)]$; and 2) Random cropping: Since the size of the base mask is larger than that of the snow-free ground truth $\mathbf{y}_i$, we randomly crop a patch of the base mask with the same size as that of $\mathbf{y}_i$ and superimpose the patch over $\mathbf{y}_i$.

We then conduct a quantitative experiment to evaluate the reliability of the synthesized snowy images $\mathbf{x}$. As previously mentioned, ground truths are the images without falling snow. To avoid potential cognitively unreasonable cases such as photography in an indoor environment or a shot captured outdoors but with the sun in the sky, we randomly select 500 semantically reasonable synthesized snowy images to conduct a fair comparison for the following survey.

To do so, 30 randomly selected snowy images are shown in each survey and we asked a subject to determine whether

[2]Snow100K dataset: https://goo.gl/BrRc3U

TABLE III

CONFUSION MATRIX OF THE SURVEY RESULTS, WHERE *Positive* AND *Negative* DENOTE THE SYNTHESIZED AND REALISTIC SNOWY IMAGES, RESPECTIVELY

| | | Prediction | | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| Ground truth | Positive | 1057 | 589 | 1646 |
| | Negative | 698 | 1166 | 1864 |
| | Total | 1755 | 1755 | |

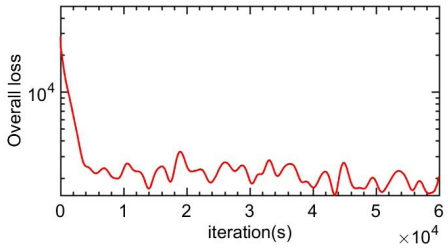



Fig. 7. Training convergence curve of $\mathcal{L}_{overall}$ as defined in Eq. (8).

each of the given samples is synthetic or realistic. Because of this random selection policy, the numbers of synthesized and realistic images in each survey could be different. In total, 117 individuals participated in our survey. As can be seen in the confusion matrix shown in Table III, our synthetics attain a recall rate as low as 64.2%, which is close to the ideal case of 50%.

## V. EXPERIMENTS

### A. Implementation Details

**Descriptor.** We remove the global pooling layer, and set the kernel size and stride of all pooling operations in Inception-v4 [29] to $3 \times 3$ and $1 \times 1$, respectively, to maintain the spatial information. The width and height of every feature maps are identical to that of the input image, and the number of kernels in each layer is set at 1/2 of those in [29] due to the limitation of GPU memory. Also, $\gamma$ as defined in Eq. (3) is set at 4, and the number of kernels in DP is set at 1/2 of the input channel.

**Recovery Submodule.** We set $\beta$ as defined in Eqs. (4) and (6) at 4 so that the kernel sizes range from $\text{conv}_{1\times1}$ to $\text{conv}_{7\times7}$. We implement the convolution kernels of sizes $\text{conv}_{5\times5}$ and $\text{conv}_{7\times7}$ in both the pyramid maxout and pyramid sum with vectors of sizes $\text{conv}_{1\times5}$ and $\text{conv}_{1\times7}$, respectively, to increase speed without reducing accuracy [29]. We use PReLU [36] as the activation function on the outputs of SE and AE.

**Training details.** The size of the training batch is set at 5, and we randomly crop a patch of size $64 \times 64$ from the samples in Snow100K's training set as the training sample. The learning rate is set at $3e^{-5}$ with the Adam Optimizer [37], and the weightings of our model are initialized by Xavier Initialization [38]. Figure 7 illustrates the training convergence of the loss defined in Eq. (8) over iterations, where $\lambda_{\hat{\mathbf{z}}} = 3$ and $\lambda_{\mathbf{w}} = 5e^{-4}$ are set in our work.

### B. Ablation Study

Tables IV-VIII illustrate the influences of different factors in the proposed DesnowNet (abbr. DSN). Two widely used metrics, PSNR and SSIM, are adopted to evaluate the snow removal quality regarding signal and structure similarities. Specifically, we evaluate the difference between the snow-free output $\hat{\mathbf{y}}$ and the corresponding ground truth $\mathbf{y}$ via these two metrics. In ablation studies, we randomly selected 2k samples from within each of the test subsets, Snow100K-S, Snow100K-M, and Snow100K-L, for evaluation, so that there are a total of 6k samples for deriving averages. Notably, only the evaluated terms are modified in the following experiments, and the remainder of the proposed DesnowNet is kept identical to the parameters introduced in Section III.

TABLE IV

COMPARISON OF DIFFERENT DESCRIPTORS IN DESNOWNET

| Metric | PSNR | SSIM |
|---|---|---|
| DSN w/ Inception-v4 [29] | 27.898 | 0.8937 |
| DSN w/ I.-v4 + ASPP [32] | 28.7691 | 0.9181 |
| DSN w/ I.-v4 + DP | 30.1741 | 0.9303 |

TABLE V

COMPARISON OF DIFFERENT NETWORKS ON SE AND AE

| Metric | PSNR | SSIM |
|---|---|---|
| DSN w/ $\text{conv}_{3\times3}$ | 28.2275 | 0.9155 |
| DSN w/ pyramid sum | 28.5529 | 0.9232 |
| DSN w/ pyramid maxout | 30.1741 | 0.9303 |

TABLE VI

COMPARISON OF DIFFERENT ARCHITECTURES OF DESNOWNET

| Metric | PSNR | SSIM |
|---|---|---|
| DSN w/o TR | 27.5136 | 0.8911 |
| DSN w/o RG | 28.0149 | 0.8974 |
| DSN w/o AE | 29.086 | 0.9088 |
| DSN | 30.1741 | 0.9303 |

TABLE VII

COMPARISON OF DIFFERENT LOSS FUNCTIONS IN DESNOWNET

| Metric | PSNR | SSIM |
|---|---|---|
| DSN w/ $l2$-norm | 29.223 | 0.9127 |
| DSN w/ pyramid $l2$-norm | 30.1741 | 0.9303 |

TABLE VIII

COMPARISON OF DIFFERENT ACTIVATION FUNCTIONS ON THE OUTPUTS OF SE AND AE

| Metric | PSNR | SSIM |
|---|---|---|
| Sigmoid | 27.8894 | 0.8985 |
| PReLU [36] | 30.1741 | 0.9303 |

**Descriptor.** Table IV shows the performances achieved using different backbones for descriptors $D_t$ and $D_r$. As can be seen, the concatenation operation utilized in the proposed dilation pyramid (DP) as defined in Eq. (3) is able to effectively aggregate the features extracted by multi-scale dilated convolutions. It also reaches a competitive PSNR and SSIM in contrast to those of JORDER [19] and DeepLab [32].

TABLE IX

PERFORMANCES OF VARIOUS METHODS ON Snow100K'S TEST SET. THE RESULTS IN *Synthesized Data* ROW DENOTE THE SIMILARITIES BETWEEN THE SYNTHESIZED SNOWY IMAGE $\mathbf{x}$ AND THE SNOW-FREE GROUND TRUTH $\mathbf{y}$; THE *Overall* COLUMN PRESENTS THE AVERAGES OVER THE ENTIRE TEST SET

| Subset | Snow100K-S | | Snow100K-M | | Snow100K-L | | Overall | |
|---|---|---|---|---|---|---|---|---|
| Metric | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Synthesized data | 25.1026 | 0.8627 | 22.8238 | 0.838 | 18.6777 | 0.7332 | 22.1876 | 0.8109 |
| Zheng *et al.* [14] | 24.3268 | 0.8108 | 22.9917 | 0.7944 | 19.9562 | 0.7221 | 22.4153 | 0.7756 |
| DerainNet [21] | 25.7457 | 0.8699 | 23.3669 | 0.8453 | 19.1831 | 0.7495 | 22.7652 | 0.8216 |
| DehazeNet [11] | 24.9605 | 0.8832 | 24.1646 | 0.8666 | 22.6175 | 0.7975 | 23.9091 | 0.8488 |
| DeepLab [32] | 25.9472 | 0.8783 | 24.3677 | 0.8572 | 21.2931 | 0.7747 | 23.8693 | 0.8367 |
| JORDER [19] | 25.6202 | 0.8861 | 24.974 | 0.8716 | 23.4085 | 0.8091 | 24.6626 | 0.8554 |
| Ours | 32.3331 | 0.95 | 30.8682 | 0.9409 | 27.1699 | 0.8983 | 30.1121 | 0.9296 |

(a) (b) (c) (d)

Fig. 8. Estimated snow-free results $\hat{\mathbf{y}}$ of the proposed DesnowNet. (a)-(b) Realistic snowy images and corresponding results, respectively. (c)-(d) Enlarged crops of (a) and (b), respectively.

In addition, although we utilized the same idea of dilated convolution [39], we also involve the typical convolution before the output layer as inspired by Yu *et al.*'s work [39] to ease the potential gridding artifact as shown in Fig. 8. Specifically, the dilation pyramid as defined Eq. (3) is equivalent to the normal convolution layer if the dilation factor $\gamma = 0$.

**Pyramid maxout.** Table V shows the performances of different networks for both SE and AE. To evaluate the influence without the pyramid maxout, we implement a basic $\text{conv}_{3\times3}$ layer to convert the dimensions of the feature map to equal those of the next layer. In addition, we change the pyramid maxout to the pyramid sum as defined in Eq. (6) in order to observe the influence. As can be seen, the pyramid sum reaches a similar performance to that of $\text{conv}_{3\times3}$ because it either cannot account for the variation in snow particle size and shape or lacks sufficient sensitivity for this task. However, this property is not necessary for predicting the snow mask $\mathbf{z}$ and chromatic aberration map $\mathbf{a}$ since these two feature maps are more concerned with accuracy rather than visual quality, which the purpose of the pyramid sum discussed in Section III-B. In contrast, the pyramid maxout yields robust prediction in this situation, with the highest PSNR and SSIM.

**Recovery module.** Certain components of the recovery module have been removed in order to ascertain their potential benefits. Results are shown in Table VI. By removing the TR module, the RG module directly maps snowy image $\mathbf{x}$ to the snow-free result $\hat{\mathbf{y}}$ without the prior of estimated snow mask $\hat{\mathbf{z}}$ and chromatic aberration map $\mathbf{a}$. Results show that this has a large impact on performance. On the other hand, while removing AE from DSN results in lower reduction of accuracy in contrast to the removal of other components, it still contributes 1.09 dB at PSNR.

**Pyramid loss function.** Table VII shows the results of different loss functions, which clearly indicate that the pyramid $l2$-norm as defined in Eq. (7) is superior to the $l2$-norm that utilizes only single scale of the feature.

**Activation function.** We also changed the activation function of the outputs of SE and AE to ascertain their influence. Table VIII illustrates the results, which show that the PReLU [36] achieves significantly better estimation accuracy.

### C. Comparison

Zheng *et al.*'s rule-based method, three learning-based atmospheric particle removal approaches, DerainNet [21] DehazeNet [11] and JORDER [19], and one semantic segmentation method, DeepLab [32], are all considered for the comparison of estimated snow-free results ($\hat{\mathbf{y}}$).

Due to the lack of source codes, we re-implemented all other methods with their default parameters to achieve the best performances. For DerainNet, the detailed portion of the snow-free image $\mathbf{y}$ was used as the ground truth for training. We implemented the DeepLab-lfov version for our comparison, and used the snow mask $\mathbf{z}$ as the ground truth for training DehazeNet and DeepLab. We use Eq. (5) with their estimated snow mask $\hat{\mathbf{z}}$ to directly recover $\hat{\mathbf{y}}$ without our RG module to match their network design. In the implementation of JORDER, the snow mask $\mathbf{z}$ and snow-free image $\mathbf{y}$ are used as its ground truth for training.

Table IX presents the quantitative results obtained via three test subsets of the Snow100K dataset. It shows that our DesnowNet outperformed other state-of-the-art methods, with JORDER achieving the second-best accuracy rates. Figure 9 shows the results of the estimated snow-free outputs $\hat{\mathbf{y}}$. Among these, DerainNet removes almost nothing from the snowy images because of its strong assumption for spatial frequency. Although Zheng *et al.*'s method [14] removes snow from snowy images, its corresponding results suffer from significant blurry artifacts and fail to preserve image details. DehazeNet is effective at removing many translucent and opaque snow particles from snowy images. However, it introduces apparent artifacts in the resultant images. Results obtained by

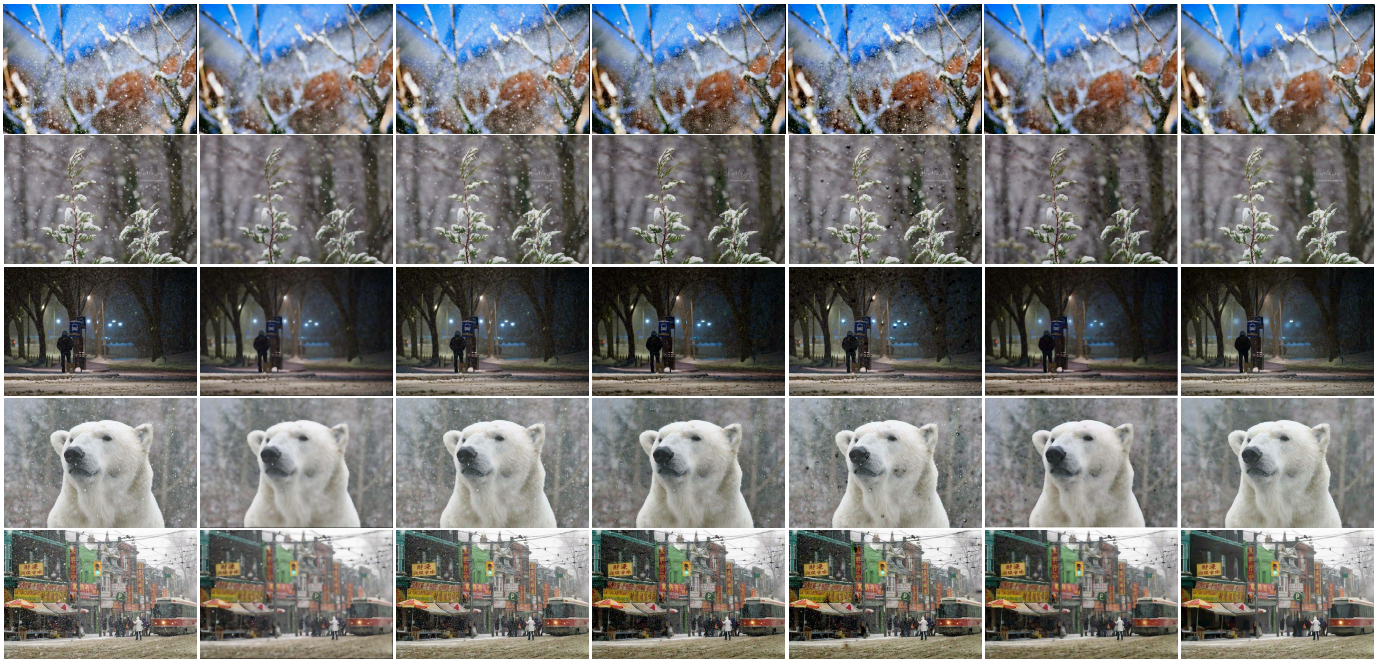



Fig. 9. Estimated snow-free results $\hat{\mathbf{y}}$ of various methods for realistic snowy images. (a) Input image. (b) Zheng *et al*. [14]. (c) DerainNet [21]. (d) DehazeNet [11]. (e) DeepLab [32]. (f) JORDER [19]. (g) Ours. Best viewed on screen.

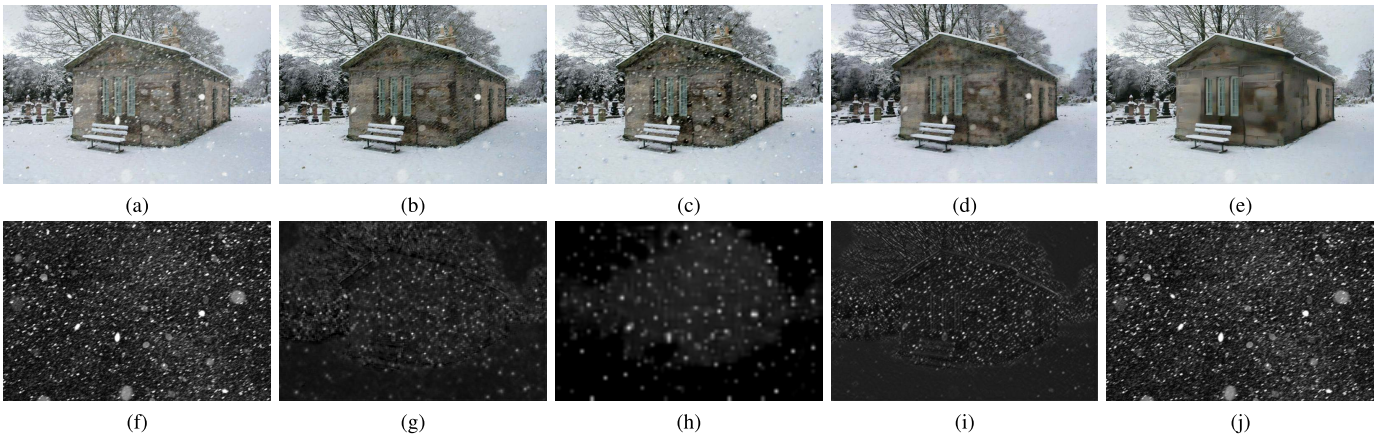



Fig. 10. Snow removal results of the proposed DesnowNet and other state-of-the-art methods. (a) Synthesized image ($\mathbf{x}$). (b) DehazeNet ($\hat{\mathbf{y}}$). (c) DeepLab ($\hat{\mathbf{y}}$). (d) JORDER ($\hat{\mathbf{y}}$). (e) Ours ($\hat{\mathbf{y}}$). (f) Synthesized snow mask ($\mathbf{z}$). (g) DehazeNet ($\hat{\mathbf{z}}$). (h) DeepLab ($\hat{\mathbf{z}}$). (i) JORDER ($\hat{\mathbf{z}}$). (j) Ours ($\hat{\mathbf{z}}$).

the DeepLab network exhibit even stronger artifacts than DehazeNet, and even introduce black spots due to the loss of spatial information. Although JORDER introduces fewer artifacts among other methods, its results still remain artifacts and snows. Conversely, experimental results show that our method introduced the fewest artifacts in its estimated snow-free outputs $\hat{\mathbf{y}}$, and that it achieves the most visually appealing results.

An additional experiment was conducted to determine the accuracy of the estimated snow mask $\hat{\mathbf{z}}$. Because of the lack of intermediate probability estimation in DerainNet and Zheng *et al*.'s method, we excluded them from this comparison. Table X presents the average similarity scores of the snow mask ground truth $\mathbf{z}$ and the estimated $\hat{\mathbf{z}}$. As can be observed, our method yields a superior similarity score from both PSNR and SSIM metrics. Of these, the large SSIM superiority particularly contrasts with other methods and is primarily due to our capability of interpreting variation in the spatial frequency, trajectory, translucency, size and shape of snow particles. The qualitative comparison shown in Fig. 10

TABLE X
ACCURACY COMPARISON OF THE ESTIMATED SNOW MASK $\hat{\mathbf{z}}$ ON SNOW100K'S TEST SET

| Metric | PSNR | SSIM |
|---|---|---|
| DehazeNet [11] | 19.622 | 0.3271 |
| DeepLab [32] | 18.8679 | 0.2942 |
| JORDER [19] | 19.9542 | 0.3917 |
| Ours | 22.0054 | 0.5662 |

TABLE XI
SUBJECTIVE EVALUATION RESULTS OF VARIOUS METHODS

| Method | Score |
|---|---|
| Zheng *et al.* [14] | 2.73 |
| DerainNet [21] | 2.21 |
| DehazeNet [11] | 3.10 |
| DeepLab [32] | 1.82 |
| JORDER [19] | 3.29 |
| Ours | 4.15 |

validates this superiority. In this experiment, DehazeNet was able to interpret only the fine-grained snow particles as shown in Fig. 10(g). While DeepLab can describe snow particles of different sizes as shown in Fig. 10(h), its down-sampling design loses much spatial information, resulting in its failure to identify small-size snow particles. Although JORDER achieves a better performance than DehazeNet and DeepLab as shown in Figs. 10 (d) and 10(i) in terms of accurately locating the snows $\hat{\mathbf{z}}$ and visual clarity of $\hat{\mathbf{y}}$, its result still suffer from artifacts and undiscovered snows due to the insufficient interpretability of the residual generation network and the lack of ability to capture all the variance of snows. Figure 10(j) presents the results of our method, which clearly shows that it successfully interpreted the variations of snow particles to yield an accurate snow-free output, as illustrated in Fig. 10(e).

We follow the identical setting in Fu *et al.*'s user study [21] to objectively evaluate the quality of snow removal results. Specifically, we showed 10 input realistic snowy images along with 50 estimated snow-free results obtained by different approaches in a random order individually to 20 subjects. Each subject was asked to score each of the estimated snow-free results from 1 to 5 regarding the image quality and the amount of visible falling snows. In total, 100 realistic snowy images were employed in this experiment. The average score of each method is shown in Table XI. As can be seen, the proposed DesnowNet achieves the highest score and DeepLab had the lowest score due to the presence of black spots in its results as shown in Fig. 9(e).

## VI. CONCLUSIONS

We have presented the first learning-based snow removal method for the removal of snow particles from a single image. Also, we have demonstrated that the baseline of a well-known semantic segmentation method and other state-of-the-art atmospheric particle removal approaches, are unable to adapt to the challenging snow removal task. We observe that the multistage network design with translucency recovery (TR) and residual generation (RG) modules successfully recovers image details obscured by opaque snow particles as well as compensates for potential artifacts. In addition, the results presented in Table VI demonstrate that the chromatic aberration map, which interprets subtle color inconsistencies of snow in three color channels as shown in Fig. 5(a), is the key milestone towards accurate prediction. Moreover, our multi-scale designs endow the network with interpretability that can account for variations of snow particles, as demonstrated in Fig. 10.

## REFERENCES


[1] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[2] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 72, Aug. 2008.
[3] Y. Wang, S. Liu, C. Chen, and B. Zeng, "A hierarchical approach for rain or snow removing in a single color image," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3936–3950, Aug. 2017.
[4] B.-H. Chen and S.-C. Huang, "Edge collapse-based dehazing algorithm for visibility restoration in real scenes," *J. Display Technol.*, vol. 12, no. 9, pp. 964–970, Sep. 2016.
[5] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
[6] B.-H. Chen, S.-C. Huang, and F.-C. Cheng, "A high-efficiency and high-speed gain intervention refinement filter for haze removal," *J. Display Technol.*, vol. 12, no. 7, pp. 753–759, Jul. 2016.
[7] S.-C. Huang, B.-H. Chen, and Y.-J. Cheng, "An efficient visibility enhancement algorithm for road scenes captured by intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2321–2332, Oct. 2014.
[8] B.-H. Chen and S.-C. Huang, "An advanced visibility restoration algorithm for single hazy images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 4, 2015, Art. no. 53.
[9] S. C. Huang, J. H. Ye, and B. H. Chen, "An advanced single-image visibility restoration algorithm for real-world hazy scenes," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 2962–2972, May 2015.
[10] S.-C. Huang, B.-H. Chen, and W.-J. Wang, "Visibility restoration of single hazy images captured in real-world weather conditions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1814–1824, Oct. 2014.
[11] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
[12] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
[13] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1742–1755, Apr. 2012.
[14] X. Zheng, Y. Liao, W. Guo, X. Fu, and X. Ding, "Single-image-based rain and snow removal using multi-guided filter," in *Proc. Int. Conf. Neural Inf. Process.*, pp. 258–265, 2013.
[15] D.-Y. Chen, C.-C. Chen, and L.-W. Kang, "Visual depth guided color image rain streaks removal using sparse coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1430–1455, Aug. 2014.
[16] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3397–3405.
[17] C.-H. Son and X.-P. Zhang, "Rain removal via shrinkage of sparse codes and learned rain dictionary," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.
[18] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2736–2744.
[19] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1357–1366.
[20] X. Fu, J. Huang, D. Z. Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1715–1723.
[21] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2944–2956, Jun. 2017.
[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 886–893.
[23] J. Bossu, N. Hautière, and J.-P. Tarel, "Rain or snow detection in image sequences through use of a histogram of orientation of streaks," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 348–367, 2011.
[24] S.-C. Pei, Y.-T. Tsai, and C.-Y. Lee, "Removing rain and snow in a single image using saturation and visibility features," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.
[25] D. Rajderkar and P. Mohod, "Removing snow from an image via image decomposition," in *Proc. Int. Conf. Emerg. Trends Comput., Commun. Nanotechnol. (ICE-CCN)*, Mar. 2013, pp. 576–579.

[26] J. Xu, W. Zhao, P. Liu, and X. Tang, "An improved guidance image based method to remove rain and snow in a single image," *Comput. Inf. Sci.*, vol. 5, no. 3, pp. 49–55, 2012.
[27] J. Xu, W. Zhao, P. Liu, and X. Tang, "Removing rain and snow in a single image using guided filter," in *Proc. IEEE Int. Conf. Comput. Sci. Autom. Eng. (CSAE)*, vol. 2. May 2012, pp. 304–307.
[28] J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. Int. Conf. Comput. Vis.-Workshop Geometry Meets Deep Learn. (ICCVW)*, vol. 3. 2017, no. 9, pp. 887–895.
[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017.
[30] A. Brock, T. Lim, J. Ritchie, and N. Weston. (2016). "Neural photo editing with introspective adversarial networks." [Online]. Available: https://arxiv.org/abs/1609.07093s
[31] S. Diamond, V. Sitzmann, S. Boyd, G. Wetzstein, and F. Heide. (2017). "Dirty pixels: Optimizing image classification architectures for raw sensor data." [Online]. Available: https://arxiv.org/abs/1701.06487
[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915
[33] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122
[34] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. (2013). "Maxout networks." [Online]. Available: https://arxiv.org/abs/1302.4389
[35] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.
[37] D. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, May 2010.
[39] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. Comput. Vis. Pattern Recognit.*, vol. 1. 2017.

**Yun-Fu Liu** (S'09–M'13) received the master's degree in electrical engineering from Chang Gung University, Taoyuan, Taiwan, in 2009, and the Ph.D. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2013.

He was involved in research with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, in 2012. In 2013, he joined the Multimedia Signal Processing Laboratory, National Taiwan University of Science and Technology, Taipei, Taiwan, as a Post-Doctoral Fellow. In 2015, he was involved in research with the Digital Video and Multimedia Laboratory, Columbia University, NY. In 2016, he joined Viscovery, Taipei, Taiwan, as a Senior Data Scientist. In 2018, he joined Alibaba DAMO Academy, Hangzhou, China, as an Algorithm Expert. His general interests lie in deep learning, computer vision, multimedia processing, and their related applications.

Dr. Liu was a recipient of the Master's Thesis Awards from the Taiwan Fuzzy Systems Association, and ChiMei Optoelectronics in 2009, the Excellent Paper Award from the Computer Vision, Graphics and Image Processing, in 2013, the Doctoral Dissertation Excellence Awards from the Taiwanese Association for Consumer Electronics, the Institute of Information and Computing Machinery, and the Image Processing and Pattern Recognition Society of Taiwan, in 2013 and 2014, and the International Computer Symposium in 2014.

**Da-Wei Jaw** received the B.S. and M.S. degrees in electronic engineering from the National Taipei University of Technology, Taipei, Taiwan, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with National Taiwan University. His research interests relating to digital image processing, machine learning, and neural networks.

**Shih-Chia Huang** is currently a Full Professor with the Department of Electronic Engineering, National Taipei University of Technology, Taiwan, and an International Adjunct Professor with the Faculty of Business and Information Technology, University of Ontario Institute of Technology, Canada. He has been named a Senior Member of the Institute of Electrical and Electronic Engineers (IEEE). He is currently the Chair of the IEEE Taipei Section Broadcast Technology Society. He was a Review Panel Member of the Small Business Innovation Research program for the Department of Economic Development, Taipei City, and New Taipei City, respectively.

He received the B.S. degree from National Taiwan Normal University, the M.S. degree from National Chiao Tung University, and the Ph.D. degree in electrical engineering from National Taiwan University, Taiwan. He has authored over 80 journal and conference papers and holds over 60 patents in the United States, Europe, Taiwan, and China. He was a recipient of the Kwoh-Ting Li Young Researcher Award in 2011 by the Taipei Chapter of the Association for Computing Machinery, Dr. Shechtman Young Researcher Award in 2012 by the National Taipei University of Technology, and the 5th National Industrial Innovation Award in 2017 by the Ministry of Economic Affairs, Taiwan. He was also the recipient of an Outstanding Research Award from National Taipei University of Technology in 2014 and the College of Electrical Engineering and Computer Science, National Taipei University of Technology from 2014 to 2016.

He is the Services and Applications Track Chair of the Applications Track Chair of the IEEE BigData Congress in 2015, a General Chair of the IEEE BigData Taipei Satellite Session from 2015 to 2016, the IEEE CloudCom conference from 2016 to 2017, and the Deep learning, Ubiquitous and Toy Computing Minitrack Chair of the Hawaii International Conference on System Sciences from 2017 to 2018. In addition, he has been an Associate Editor of the *Journal of Artificial Intelligence* and a Guest Editor of the *Information Systems Frontiers* and the *International Journal of Web Services Research*.

His research interests include intelligent multimedia systems, image processing and video coding, video surveillance systems, cloud computing and big data analytics, artificial intelligence, and mobile applications and systems.

**Jenq-Neng Hwang** (F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, the Ph.D. degree from the University of Southern California. In 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, where he has been a Full Professor since 1999. He served as the Associate Chair for research from 2003 to 2005 and from 2011 to 2015. He is currently an Associate Chair for global affairs and international development with the EE Department. He has authored over 330 journal, conference papers and book chapters in the areas of machine learning, multimedia signal processing, and multimedia system integration and networking, including an authored textbook on *Multimedia Networking: from Theory to Practice*, (Cambridge University Press). He has close working relationship with the industry on multimedia signal processing and multimedia networking.

Dr. Hwang was the Society's representative to IEEE Neural Network Council from 1996 to 2000. He is a founding member of Multimedia Signal Processing Technical Committee of IEEE Signal Processing Society. He is currently a member of Multimedia Technical Committee of IEEE Communication Society and also a member of Multimedia Signal Processing Technical Committee of IEEE Signal Processing Society. He was a recipient of the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He served as the Program Co-Chair of the IEEE ICME 2016 and was the Program Co-Chairs of ICASSP 1998 and ISCAS 2009. He served as an Associate Editors for the IEEE T-SP, T-NN, and T-CSVT, T-IP and *Signal Processing Magazine*. He is currently on the Editorial Board of ZTE Communications, ETRI, IJDMB, and JSPS journals.