Learning Multiple Adverse Weather Removal via Two-stage Knowledge Learning and Multi-contrastive Regularization: Toward a Unified Model

Wei-Ting Chen[1*], Zhi-Kai Huang[2*],
Cheng-Che Tsai[1], Hao-Hsiang Yang[2], Jian-Jiun Ding[2], and Sy-Yen Kuo[2]

[1]Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taiwan

{f05943089, r10921059, r08943148, r05921014, jjding, sykuo}@ntu.edu.tw

Abstract

In this paper, an ill-posed problem of multiple adverse weather removal is investigated. Our goal is to train a model with a 'unified' architecture and only one set of pre-trained weights that can tackle multiple types of adverse weathers such as haze, snow, and rain simultaneously. To this end, a two-stage knowledge learning mechanism including knowledge collation (KC) and knowledge examination (KE) based on a multi-teacher and student architecture is proposed. At the KC, the student network aims to learn the comprehensive bad weather removal problem from multiple well-trained teacher networks where each of them is specialized in a specific bad weather removal problem. To accomplish this process, a novel collaborative knowledge transfer is proposed. At the KE, the student model is trained without the teacher networks and examined by challenging pixel loss derived by the ground truth. Moreover, to improve the performance of our training framework, a novel loss function called multi-contrastive knowledge regularization (MCR) loss is proposed. Experiments on several datasets show that our student model can achieve promising results on different bad weather removal tasks simultaneously. The code is available in our project page.

1. Introduction

Adverse weather such as haze, rain, snow, and adherent raindrop is a common phenomenon in our daily life. It may usually degrade the visibility of images and deteriorate the performance of high-level vision applications (e.g., object detection and semantic segmentation). To tackle this problem, several adverse weather restoration algorithms such as deraining [1–10], desnowing [11–13], dehazing [14–20], and all in one bad weather removal [21] have been widely explored in past decades. Although these methods achieve a promising performance, there still exists a limitation for deploying adverse weather removal in real-world applications such as surveillance systems, autonomous vehicle systems, or edge devices due to the high extension cost. Specifically, existing approaches cannot address several weather types in a unified architecture or a set of pre-trained weights simultaneously. In real-world scenarios, it is unavoidable to handle various weather types. As shown in Fig. 1, the existing methods may have several limitations and we summarize them as follows.

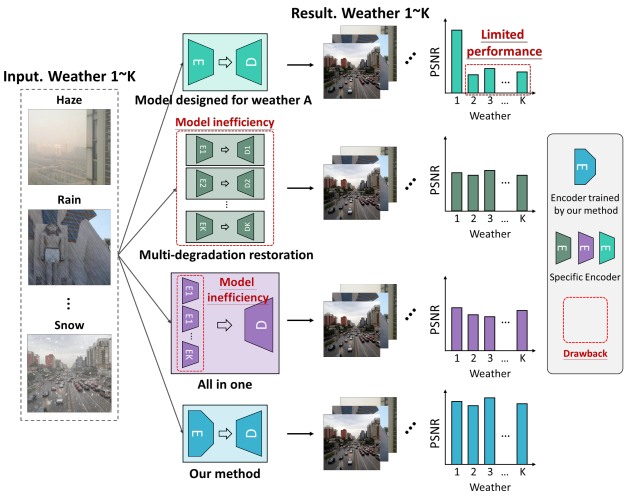


Figure 1. **Overview of the existing bad weather removal algorithms.** Our proposed method can achieve promising performance in comprehensive bad weather removal problems without additional cost at the inference stage.

(i) Single Weather Removal Algorithms: For most single weather removal algorithms [15, 22, 23], although they can achieve promising results in the specific weather, they

*Indicates equal contribution.

may have limited performance on other types of weather because the features of them are not considered. Thus, in real-world applications, the systems need to determine the weather type first and then select a corresponding adverse weather restoration approach.

(ii) Multi-degradation Removal Algorithms: Some recent studies aim to tackle multiple degradation problems [18, 24–26] via a single framework. Nevertheless, they generally require several sets of pre-trained weights for various degradations. It requires the network to adopt different pre-trained weights according to weather types, which is troublesome and inefficient.

(iii) All-in-one Bad Weather Removal methods: In recent years, the all-in-one bad weather removal model [21] has attracted considerable attention because it can handle several types of weather in a set of pre-trained weights by using the neural architecture search (NAS) technique. Although this method can achieve encouraging performance on various types of bad weather, it has the model inefficiency problem. Specifically, the model size of this method may increase dramatically if the model needs to solve more types of weather since more feature extractors are required.

For a real-world outdoor system, the restoration model should be able to be extended to other weather types without additional cost while can achieve decent reconstruction performance. To achieve this goal, inspired by knowledge distillation [27], we proposed a novel method for the adverse weather removal based on a two-stage knowledge learning process including knowledge collation (KC) and knowledge examination (KE). For the former stage, several well-trained teacher models guide the 'immature' student model to integrate and learn the knowledge of various weather types by the proposed collaborative knowledge transfer (CKT) technique. The CKT consists of progressive feature projector and bi-directional feature matching to tackle the knowledge transfer for multiple adverse weather removal networks. These two mechanisms can constrain and improve the feature learning process in the common feature space. For the latter stage, the goal is to improve the robustness of the 'mature' student network for comprehensive weather removal by examining it with challenging constraints. Moreover, to enhance the robustness, multi-contrastive regularization (MCR) is developed to optimize the student network by improving its discriminative ability for different weather types.

The contributions of this paper are summarized as:

- A novel method for the comprehensive bad weather removal based on two-stage knowledge learning is proposed. During the test stage, the network can tackle different weather removal problems with a unified architecture and one set of pre-trained parameters.
- To boost the performance of the proposed training scheme, the CKT and MCR are designed.
- Extensive experiments are conducted to verify that the proposed training scheme can achieve promising results on several adverse weather types simultaneously.

2. Related Works

2.1. Adverse Weather Removal

There are several image restoration algorithms for adverse weather, including deraining [1–5, 7, 9, 10, 28–31], dehazing/defogging [14, 15, 32–36], desnowing [37–42], multi-degradation removal [18, 24–26, 43], and all in one strategy [21].

Single Weather Removal. We briefly introduce different single weather removal methods. *For haze removal*, Qu *et al.* [16] proposed a GAN-based enhanced Pix2pix network to generate haze-free images. Dong *et al.* [17] developed dense feature fusion to reconstruct the missing spatial information. Wu *et al.* [44] adopted a contrastive regularization learning technique and a dynamic feature enhancement module for haze removal. *For rain removal*, Li *et al.* [45] adopted the recurrent network to capture rain streak information. Yang *et al.* [28] developed the deep neural network to learn the intensity and the location of rain streaks jointly. Deng *et al.* [8] proposed DRD-Net which consists of a rain residual network and a detailed repair network. Quan *et al.* [46] proposed an architecture based on NAS to handle both rain streaks and raindrops simultaneously. *For snow removal*, Liu *et al.* [13] adopted the Inception-v4 model to construct a two-stage snow removal network termed DesnowNet. Chen *et al.* [11] proposed a joint size and transparency snow removal process to tackle snow particles with non-transparency and various sizes. Jaw *et al.* [47] proposed to combine high-level semantic features and other feature maps to handle snow removal problem. Chen *et al.* [12] introduced the dual-tree wavelet transformation to a network dubbed HDCW-Net for snow information retrieval.

Although the aforementioned works achieve promising performance on a specific weather type, they may not generate decent results on other types of adverse weather.

Multi-degradation removal. Zou *et al.* [24] proposed a unified framework which consists of a discriminative network called "Separation-Critic" and a crossroad L_1 loss function. Zamir *et al.* [18] proposed a multi-stage strategy image restoration network called MPRNet which adopts the attention module to refine the incoming features at each stage. Pan *et al.* [25] proposed a general architecture that focuses on estimating structures and details simultaneously in parallel branches.

Though these strategies can obtain encouraging results in various weather types with a unified framework, they require several sets of pre-trained weights to deal with different weather types.

All-in-one Bad Weather Removal. Li *et al.* [21] proposed an end-to-end training scheme based on the NAS architecture to search crucial features from multiple encoders for different weather types. Then, the reconstructed images are optimized by categorical adversarial learning to generate a robust network for various weather types.

Though this method can achieve encouraging results in several weather types, the model size increases rapidly when it needs to handle more weather types since each of them requires its own encoder.

Knowledge Distillation and Contrastive Learning. Knowledge distillation (KD) [27] is to transfer the knowledge of a large teacher model to a smaller student network via a teacher-student architecture. This idea was extended by [48, 49] which used the intermediate representation extracted from the teacher to assist the training process of the student network. The KD model achieves encouraging results in several topics such as object detection [50], semantic segmentation [51], and image restoration [33].

Contrastive learning has attracted great attention in several computer vision tasks such as image retrieval [52], ReID [53], image classification [54], and face recognition [55]. Its key idea is to make positive samples attractive and negative samples repelled by the contrastive loss [56].

3. Proposed Method

3.1. Problem Formulation

In this work, we aim to tackle the multiple adverse weather removal problem via a unified architecture and a set of pre-trained weights. Inspired by the KD [27], we proposed a training scheme that can transfer the knowledge collaboratively to the student model from multiple teacher models which are specialized for various weather removal problems. Its detail is as follows. Given K well-trained teacher models $\{T_i\}_{i=1}^{K}$, each of them conducts a specific weather removal task. Let W_i be the type of weather tackled by model T_i. We further assume that $W_i \neq W_j, \forall i \neq j$ to prevent the model from losing generality. The proposed method aims to train a compact model which can address the comprehensive bad weather removal problem in a unified architecture at the testing stage. That is, this model can remove K types of bad weathers in $W = \bigcup_{i=1}^{K} W_i$. This task is challenging because the model should contain the knowledge for several weather types simultaneously. Moreover, the performance can be maintained without additional model costs. We illustrate the proposed strategy in the following subsections.

3.2. Two-stage Knowledge Learning

Existing single weather removal algorithms based on KD [33] usually leverage a teacher network trained with clear images and transfer the knowledge to a student network trained on single weather degradation. However, the performance may be limited under the multi-weather scenarios since the student network does not learn about discriminative features for different weather types due to the lack of appropriate guidance. To address this issue, as shown in Fig. 2, we proposed a two-stage knowledge learning scheme which is illustrated as follows.

Knowledge Collation (KC). In KC, there are several well-trained teacher networks and one student network. Each teacher network is specialized for one weather removal and the student network aims to *learn* and *collate* the knowledge from the teachers to achieve comprehensive weather removal. In each epoch, the student network is trained with different teacher networks concurrently. Since the student network is not capable of mimicking the perfect representations from the ground truth at this stage, we proposed to conduct easier regularization, that is, the loss calculation is based on the results predicted by the teacher networks. Moreover, to accomplish the knowledge transfer robustly, we proposed a novel technique called collaborative knowledge transfer which is presented in subsection 3.3.

Knowledge Examination (KE). After the KC, we can assume the student network is 'mature' and can achieve promising results in different weather types. Thus, this stage aims to strengthen the robustness and the discriminative ability of the network via examining the student network by more demanding constraints. To this end, the student network is trained without the guidance of teacher networks and the more challenging regularization is applied.

3.3. Collaborative Knowledge Transfer

The architecture of the proposed collaborative knowledge transfer (CKT) is presented in Fig. 3. The features generated by several teacher networks and a student network are projected to a common feature space via the progressive feature projector (PFP). Then, to achieve robust and effective knowledge transfer, bi-directional feature matching (BFM) is conducted. The details are illustrated as follows.

Progressive Feature Projector. For most KD-based strategies [33], the knowledge transfer process usually applies a feature adaptor module to project the features of the teacher network to the feature space of the student for the feature alignment. This strategy may be effective for those tasks because they only involve the single knowledge transfer (e.g., transferring the knowledge of clear features to the student network). However, it may not be appropriate for our scenario because the transferred knowledge is more diverse (i.e., multi-weather networks). The features projected from teacher networks cannot guarantee the feature space of the student network is optimal for the feature learning process due to the domain discrepancy of various weather

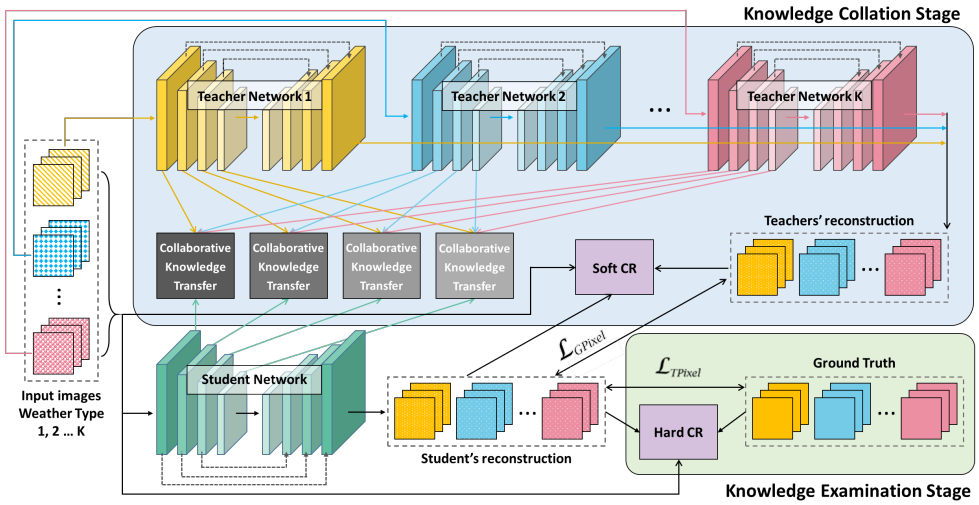


Figure 2. **The architecture of the proposed method for adverse weather removal.** It consists of two stages: KC and KE. In KC, the student network is trained with several teacher networks and a student network by projecting their features for common feature learning by the CKT. In KE, the student network are trained without the guidance of the teacher networks.

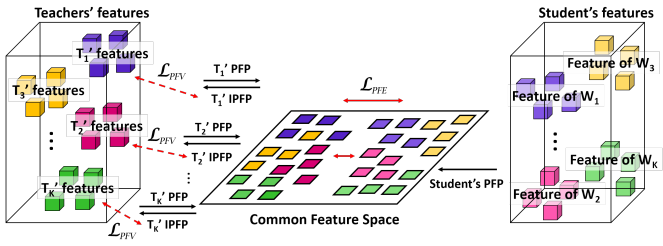


Figure 3. **The details of the proposed collaborative knowledge transfer.** $\mathcal{L}_{PFE}$ forces the learned feature from the student network to be close to that of the teacher network while $\mathcal{L}_{PFV}$ maintains the validness of the projected features.

types. Thus, to cope with this issue, as shown in Fig. 3, we proposed the progressive feature projector (PFP) which projects the features derived from *both* teacher and student networks to a common feature space. The PFP is a learnable module that can make the network determine the most suitable feature space for common feature learning. The PFP is a small network constituted by several convolution blocks with stride rate 1 and the ReLU activation function.

Then, the projected feature error ($\mathcal{L}_{PFE}$) is proposed to constrain the feature learning process. We make the projected features closer in the common feature space. We adopt the L_1 loss to calculate the distance between two projected features. Moreover, the pyramid pooling [57] is adopted to the projected features for expanding the contextual information in different levels [58]. The operation is:

$$\mathcal{L}_{PFE} = \sum_{q=1}^{Q} ||\varphi(F_{T_i}^q) - \varphi(F_S^q)||_1, \tag{1}$$

where Q denotes the total number of the layers in the encoder. $F_{T_i}^q$ presents the feature map projected by the PFP from the q^{th} layer of the encoder in the corresponding teacher network T_i. F_S^q represents the feature map projected by the PFP from the q^{th} layer of the encoder in the student network. $\varphi(\cdot)$ denotes the pyramid pooling operation.

Bidirectional Feature Matching. To learn the knowledge from several teacher networks robustly, we proposed bidirectional feature matching (BFM) to constrain the learned features. First, the projected features of teacher networks are projected back to the original input space via the inverse progressive feature projector (IPFP). Then, we calculate their difference with the original features by the loss of projected feature verification ($\mathcal{L}_{PFV}$).

$$\mathcal{L}_{PFV} = \sum_{q=1}^{Q} ||\rho(F_{T_i}^q) - \hat{F}_{T_i}^q||_1, \tag{2}$$

where $\hat{F}_{T_i}^q$ denotes the original feature without using the PFP operation. $\rho(\cdot)$ denotes the IPFP where its architecture is similar to that of the PFP. Our idea is that, the original feature generated by the teacher and the student networks may be projected to unreasonable features to minimize the $\mathcal{L}_{PFE}$. By applying this auxiliary process, the validness of the projected feature can be guaranteed and the robustness of the whole CKT process can be improved.

The total loss of the CKT ($\mathcal{L}_{CKT}$) is defined as:

$$\mathcal{L}_{CKT} = \mathcal{L}_{PFE} + \mathcal{L}_{PFV}, \tag{3}$$

3.4. Multi-contrastive Regularization

To improve the performance of the proposed training scheme, inspired by contrastive learning, we proposed multi-contrastive regularization (MCR) which is embedded in the two-stage knowledge learning process according to the ability of the network. There are two losses in MCR: the soft contrastive regularization (SCR) and the hard contrastive regularization (HCR). The former aims to improve the performance of a specific weather type while the latter enhance the discriminative ability of the network for multiple kinds of weather. We first introduce the contrastive regularization loss which is:

$$\mathcal{Q}(v, v^+, v^-) = \\ -\log\left[\frac{\exp(\Psi(v)\cdot\Psi(v^+)/\tau)}{\exp(\Psi(v)\cdot\Psi(v^+)/\tau)+\sum_{r=1}^{R}\exp(\Psi(v)\cdot\Psi(v_r^-)/\tau)}\right], \tag{4}$$

where v, v^+, and v^- denote the predicted result, positive sample, and negative sample, respectively. $\cdot$ represents the dot product operation. $\Psi(\cdot)$ is the feature extraction operation by the VGG-19 network. τ is the scale temperature which is set to 0.07 in the paper and R denotes the total number of negative samples. Then, we illustrate the proposed SCR and HCR.

Soft Contrastive Regularization. The soft contrastive regularization (SCR) is adopted at the KC stage to optimize the student network. Since the student network is not robust to mimic the behavior of the teacher network perfectly, we reduce the difficulty of regularization for the learning process. Our idea is that, the existing methods adopted contrastive learning usually directly use the ground truth of the degraded image as the positive sample. However, it is challenging for an 'immature' student network to learn such challenging sample. Thus, as shown in the left side of Fig. 4, to reduce the difficulty of the learning process, we adopt the result predicted by the teacher network as the positive sample. It enables the network to learn the feature representation easily. For the negative samples, we adopt a set of images degraded by a specific weather since we desire the network to focus on the images with the same degradation. Given an input image degraded by the weather type W_i, the SCR loss $\mathcal{L}_{SCR}$ is defined as:

$$\mathcal{L}_{SCR} = Q(\hat{J}_S, \hat{J}_{T_i}, \{\mathbf{I}_r^{W_i}\}_{r=1}^{R}), \tag{5}$$

where $\{\mathbf{I}_r^{W_i}\}_{r=1}^{R}$ is the image set degraded by the weather type W_i. $\hat{J}_S$ and $\hat{J}_{T_i}$ are the results predicted by the student network S and the teacher network T_i, respectively.

Hard Contrastive Regularization. The hard contrastive regularization (HCR) is applied at the KE stage. At this stage, we assume that the student network has been trained by the various teachers for several epochs and is 'mature' enough to handle different types of bad weather. Therefore, as shown in the right side of Fig. 4, we leverage the ground truth of the input image as the positive sample while the set of images degraded by all weather types as the negative samples. The positive sample can enforce the network to learn more accurate results while the negative samples can enhance the discriminative ability for various weather types. This operation allows the network to learn more comprehensive information between different weathers and improve its robustness to multiple weather types. The HCR loss is defined as:

$$\mathcal{L}_{HCR} = Q(\hat{J}_S, J_{GT}, \{\{\mathbf{I}_r^{W_i}\}_{r=1}^{R}\}_{i=1}^{K}), \tag{6}$$

where J_{GT} is the ground truth of the input image and K is the total number of weather types.

3.5. Overall Loss

The losses of the two stages are presented as follows.

Knowledge Collation. The $\mathcal{L}_{CKT}$ is adopted on the common feature space for multi-teacher knowledge transfer. Moreover, similar to the SCR, we leverage the result predicted by the teacher network to calculate the difference since the student network is not stable and immature at this stage. The total loss of the KC stage ($\mathcal{L}_{KC}$) is:

$$\mathcal{L}_{KC} = \mathcal{L}_{TPixel} + \lambda_1\mathcal{L}_{CKT} + \lambda_2\mathcal{L}_{SCR}, \tag{7}$$

where $\mathcal{L}_{TPixel}$ is the L_1 norm of the difference between the result recovered by the teacher network and the predicted result.

Knowledge Examination. The total loss function at KE stage $\mathcal{L}_{KE}$ can be illustrated as:

$$\mathcal{L}_{KE} = \mathcal{L}_{GPixel} + \lambda_3\mathcal{L}_{HCR}, \tag{8}$$

where $\mathcal{L}_{GPixel}$ denotes the difference between the ground truth and the recovered result in terms of the L_1 norm. Similar to HCR, the capability of the student network at the KE stage is mature enough to tackle different weather types. Thus, adopting the ground truth to calculate the loss enables the network to learn the more accurate details of the result.

4. Implementation

4.1. Datasets

Various adverse weather datasets are applied including "RESIDE" [59], "Rain 1400" [2], and "CSD" [12]. "RESIDE" is a large haze dataset that consists of the "ITS" dataset and "OTS" dataset for training and the "SOTS" dataset for testing. "Rain 1400" contains 12600 synthesized rain images. "CSD" contains 10K synthesized snow images. At the training stage, we sample 5000 images from "OTS", "Rain 1400", and "CSD" as three individual training sets, respectively. We merge them as a "mixed training set".

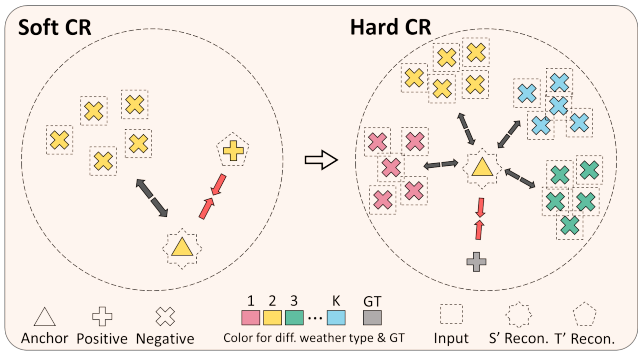


Figure 4. **Illustration of the proposed multi-contrastive regularization (MCR).**

For the testing stage, we evaluate our method on the SOTS dataset and the testing sets of "Rain 1400" and "CSD". We merge three test sets as a "mixed testing set".

4.2. Training Details

For the training detail, we need to train several teacher networks and a student network. Each of them is trained with 250 epochs and the warm-up strategy. The learning rate is 2×10^{-4}. The Adam [60] optimizer is applied. The batch size is 32 and we randomly crop all input images to 224×224. The total number of parameters is 2.8×10^7 and our method takes 16.6 ms to recover an input image. The proposed network was trained on an Nvidia Tesla V100 GPU and implemented on the Pytorch platform. In each epoch, we split 30% of training data as the validation set. We adopt the similar architecture proposed in MSBDN [17] as the backbone in our network. The scaling factors λ_1 to λ_3 are 1, 0.1, and 0.2, respectively.

Initially, the training sets for different weather types are adopted to train the corresponding teacher networks. Then, we adopt the mixed training set to train the student network. At the KC stage, we fixed the teacher networks to train the student network for 125 epochs. Then, we train the student network without the guidance of teacher networks for another 125 epochs at the KE stage.

5. Experiments

In this section, we evaluate the proposed method on both synthetic and real-world adverse weather images including haze, rain, snow. For dehazing, we compare our method with state-of-the-art approaches including the EPDN [16], the AECR-Net [44], the MSBDN [17], the PFDN [61], the KDDN [33], the FFA-Net [36]. For snow removal, we compare our method with existing desnowing approaches including the DesnowNet [13], the DesnowGAN [47], the JSTASR [11], and the HDCW-Net [12]. For the rain removal, we compare our method with the DRD-Net [8], the PReNet [62], the JORDER [63], the MSPFN [7], the DualGCN [64], and the JRGR [30]. We also adopt multi-degradation restoration methods including the DAD [24] and the MPRNet [18], and the all-in-one bad weather removal strategy [21].

5.1. Quantitative Evaluation

For quantitative evaluation, we apply the structural similarity (SSIM) and the peak signal to noise ratio (PSNR). For the single degradation and multiple degradation restoration models, two types of results are reported: (i) the model trained on specific weather (i.e., single weather training set) and (ii) the model trained on data of all weather types (i.e., the mixed training set). For all-in-one strategy [21] [1] and our method, we trained them by using the mixed training set. For a fair comparison, we retrain each compared model (if the original training code is provided) based on our training dataset and report the best result. The results are presented in Table 1. One can see that, each method has good performance when it is trained on single weather type while it may be deteriorated when it is trained on multi-weather types. From the perspective of removing single weather degradation, our method may not be the best method within each weather type. However, our method can achieve superior performance compared to other existing methods when we address all weather types by solely adopting a set of pre-trained parameters and a unified architecture.

5.2. Qualitative Evaluation

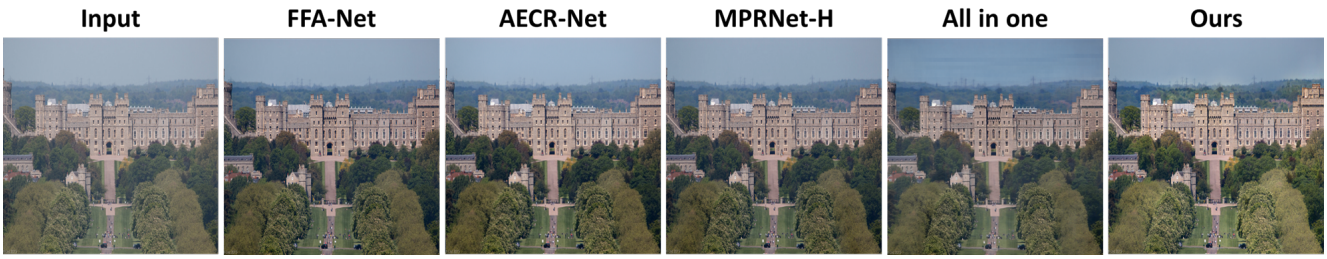
We present the visual results recovered by the proposed method in Fig. 5 under haze, rain, and snow scenarios. One can notice that our method can achieve encouraging results in visual quality in each whether type. For the hazy scenario, the result recovered by our method contains less residual haze. For snow and rain scenarios, our method can remove more snow particles and rain streaks compared with other methods.

5.3. Ablation Study

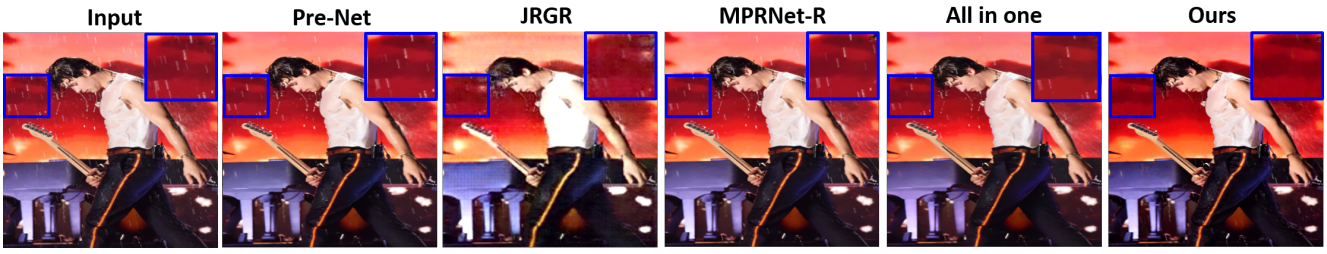
We evaluate the effectiveness of each proposed module including the collaborative knowledge transfer (CKT), the multi-contrastive regularization (MCR), and the two-stage knowledge learning strategy. We report the result tested on the mixed testing set and trained on the mixed training set.

Effectiveness of CKT. Six combinations are conducted for comparison. (Baseline): the backbone; (C1): the baseline learned the knowledge from the **S**ingle **T**eacher network trained on **C**lear images (STC) (i.e., the input and output of the teacher network are the clear images, which is similar to [33]); (C2): the baseline learned the knowl-

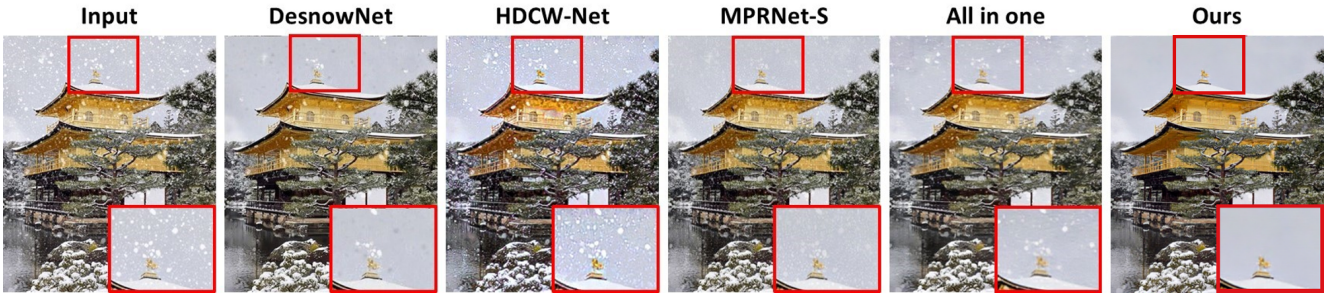
[1] Since the original codes of [21] and [13] are not available, the results in this paper are based on our implementation.



(a) **Visual comparison on dehazed images.** Our method can generate the result with less residual haze.



(b) **Rain removal results of our method compared with state-of-the-art rain removal methods.**



(c) **Comparison of desnowing results.** Our method can remove more snow particles compared with other methods.

Figure 5. **Comparison between adverse weather removal algorithms.**

edge from **S**ingle **T**eacher network trained on **M**ixed training set (STM); (C3): the proposed multi-teacher strategy where only the features of teacher networks are projected to the student's feature space. (C4): C3 with the PFP; (C5): C4 with the BFM. Note that, apart from the teacher network of 'C1', each network is trained on the mixed training set. One can see that, the student model trained by the single teacher network may have limited performance because single knowledge transfer cannot address the diverse features appropriately. Particularly, the teacher network trained on clear images has worse performance compared to C2 since it does not learn the discriminative ability for multiple weather types. Moreover, from C3 to C5, each module proposed in the CKT can contribute to the performance of the network. However, only adopting the PFP module may deteriorate the performance since the projected feature may not be valid for feature learning in the common feature space. The PFPs of the teacher and the student networks may tend to project unreasonable features to reduce the $\mathcal{L}_{PFE}$. The BFM can be applied to alleviate this issue.

Effectiveness of MCR. We further verify the effectiveness of the proposed MCR in Table 3. It is worth noting that by using the SCR and HCR at KC and KE stages respectively, the best performance can be obtained.

Effectiveness of Two-stage Training Strategy. We compare the proposed two-stage knowledge learning strategy with single stage training strategy in Table 4. Note that the single stage training strategy presents that we adopt multiple teacher networks to train student with $\mathcal{L}_{KC}$ and $\mathcal{L}_{KE}$ in the whole training process. One can see that, the two-stage training strategy achieves better performance.

Effectiveness of Multi-weather Removal. In summary, with the proposed techniques, the performance of multi-weather removal can be much improved compared to the baseline. Specifically, the values of PSNR and SSIM can be increased from 29.274 and 0.918 to 32.814 and 0.955, respectively. Moreover, to verify the effectiveness of feature extraction by the proposed approach, we present the t-SNE of the extracted features in Fig. 6a and Fig. 6b by the C1 module and our method. The results indicate that when using the existing single teacher strategy, the features extracted from the different weather types may be confused

Table 1. **Quantitative evaluation on adverse weather removal.** For the MPRNet and DAD, the suffices 'H', 'R', and 'S' denote that the model is trained on haze, rain and snow datasets, respectively. For the regions with the red color, they denote the results only trained on the model's original weather (e.g., PFDN is trained on haze dataset) and tested on the corresponding testing set. For the regions with the gray color, they denote the results trained on the mixed training set and test on three testing sets, respectively. The words with **boldface** indicate the best results in the corresponding weather.

Methods		Datasets			
		Original Weather	Haze	Rain	Snow
Dehaze	EPDN	23.82/0.89	23.18/0.87	22.20/0.76	20.16/0.77
	PFDN	31.45/0.97	27.41/0.95	31.03/0.87	27.41/0.89
	KDDN	33.49/0.97	29.16/0.94	23.36/0.87	26.15/0.87
	MSBDN	33.79/0.98	30.05/0.96	29.62/0.89	28.15/0.91
	FFA-Net	34.98/0.99	31.63/0.96	31.77/0.91	29.27/0.94
	AECRNet	**35.61/0.98**	32.26/0.97	30.43/0.91	27.07/0.92
	DAD-H	26.97/0.95	-	-	-
	MPRNet-H	31.31/0.97	-	-	-
Derain	JORDER	31.28/0.92	21.63/0.85	30.03/0.88	21.04/0.80
	PReNet	31.88/0.93	23.37/0.93	29.65/0.91	23.61/0.90
	DRD-Net	29.65/0.88	21.60/0.86	25.98/0.82	22.03/0.79
	MSPFN	29.24/0.88	24.94/0.93	27.24/0.82	20.59/0.76
	DualGCN	30.50/0.91	19.43/0.84	21.15/0.68	18.70/0.75
	JRJG	31.18/0.91	30.51/0.91	28.92/0.89	28.48/0.86
	DAD-R	31.74/0.93	-	-	-
	MPRNet-R	**33.52/0.93**	-	-	-
Desnow	DesnowNet	25.63/0.88	24.07/0.87	27.58/0.86	24.18/0.85
	JSTASR	27.52/0.87	25.65/0.85	25.51/0.81	26.03/0.84
	DesnowGAN	28.63/0.90	25.77/0.90	28.42/0.87	27.09/0.88
	HDCW-Net	29.11/0.91	30.07/0.93	27.20/0.85	28.85/0.89
	DAD-S	29.29/0.90	-	-	-
	MPRNet-S	**31.53/0.96**	-	-	-
DAD		-	25.94/0.94	29.87/0.87	26.79/0.87
MPRNet		-	29.38/0.95	31.36/0.91	29.68/0.94
All in one		-	30.49/0.95	30.82/0.90	28.65/0.92
Ours		-	**33.95/0.98**	**33.13/0.93**	**31.35/0.95**

Table 2. **Ablation study of the proposed collaborative knowledge transfer on three weather types.** Note that, since C1 requires the clean images to train its teacher network, we adopt the ground truths of the mixed training set.

Combination	Module					Metric
	STC	STM	MT	PFP	BFM	PSNR/SSIM
Baseline	-	-	-	-	-	29.274 / 0.918
C1	√	-	-	-	-	26.975 / 0.897
C2	-	√	-	-	-	29.666 / 0.921
C3	-	-	√	-	-	30.507 / 0.931
C4	-	-	√	√	-	29.784 / 0.923
C5	-	-	√	√	√	**31.668 / 0.943**

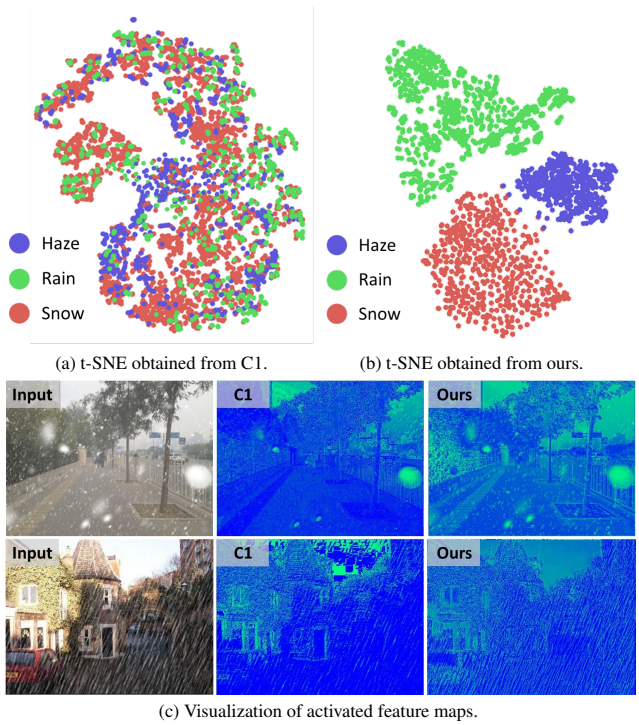
Table 3. **Ablation study of the proposed multi-contrastive regularization loss.**

Stage	KC		KE		PSNR/SSIM
Loss	$\mathcal{L}_{SCR}$	$\mathcal{L}_{HCR}$	$\mathcal{L}_{SCR}$	$\mathcal{L}_{HCR}$	
C5	-	-	-	-	31.668 / 0.943
C6	√	-	√	-	32.183 / 0.945
C7	-	√	-	√	32.264 / 0.946
C8	√	-	-	√	**32.814 / 0.955**

since the student network is not guided in a discriminative way under the multi-weather scenarios. By contrast, the proposed method can well distinguish the two types of

Table 4. **Effectiveness of the two-stage training strategy.**

Strategy	PSNR	SSIM
w/o Two-stage knowledge learning	32.061	0.945
Two-stage knowledge learning	**32.814**	**0.955**



(a) t-SNE obtained from C1. (b) t-SNE obtained from ours.

(c) Visualization of activated feature maps.

Figure 6. **Visual comparison of t-SNE and activation feature maps.**

weather and restore these degradations effectively. Moreover, we present the activated feature map in Fig. 6c and it indicates that the student model trained by our method can capture the degradation (i.e., rain or snow) of input images much more accurately.

6. Conclusion

In this paper, we proposed a novel approach to address the bad weather removal problem with a unified architecture and a set of pre-trained weights. We designed several mechanisms including two-stage knowledge learning, CKT, and MCR. The experimental results show that the proposed method can achieve encouraging performance compared to existing methods and the ablation studies prove the effectiveness of each proposed module.

7. Acknowledgement

We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, Liang-Jian Deng, and Yao Wang. A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors. In *CVPR*, 2017. 1, 2

[2] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 1, 2, 5

[3] Wenhan Yang, Jiaying Liu, Shuai Yang, and Zongming Guo. Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. *IEEE TIP*, 2019. 1, 2

[4] Wei Wei, Lixuan Yi, Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Should we encode rain streaks in video as deterministic or stochastic? In *ICCV*, 2017. 1, 2

[5] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, 2019. 1, 2

[6] Rajeev Yasarla, Vishwanath A Sindagi, and Vishal M Patel. Syn2real transfer learning for image deraining using gaussian processes. In *CVPR*, 2020. 1

[7] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020. 1, 2, 6

[8] Sen Deng, Mingqiang Wei, Jun Wang, Yidan Feng, Luming Liang, Haoran Xie, Fu Lee Wang, and Meng Wang. Detail-recovery image deraining via context aggregation networks. In *CVPR*, 2020. 1, 2, 6

[9] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *CVPR*, 2020. 1, 2

[10] Yingjun Du, Jun Xu, Qiang Qiu, Xiantong Zhen, and Lei Zhang. Variational image deraining. In *WACV*, 2020. 1, 2

[11] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Chen-Che Tsai, and Sy-Yen Kuo. Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *ECCV*, 2020. 1, 2, 6

[12] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *ICCV*, 2021. 1, 2, 5, 6

[13] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE TIP*, 2018. 1, 2, 6

[14] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *CVPR*, 2016. 1, 2

[15] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018. 1, 2

[16] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *CVPR*, 2019. 1, 2, 6

[17] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, 2020. 1, 2, 6

[18] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 2, 6

[19] Hao-Hsiang Yang and Yanwei Fu. Wavelet u-net and the chromatic adaptation transform for single image dehazing. In *ICIP*, 2019. 1

[20] Hao-Hsiang Yang, Chao-Han Huck Yang, and Yi-Chang James Tsai. Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing. In *ICASSP*, 2020. 1

[21] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, 2020. 1, 2, 3, 6

[22] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *CVPR*, 2019. 1

[23] Wei-Ting Chen, Jian-Jiun Ding, and Sy-Yen Kuo. Pms-net: Robust haze removal based on patch map for single images. In *CVPR*, pages 11681–11689, 2019. 1

[24] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *CVPR*, 2020. 2, 6

[25] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, et al. Learning dual convolutional neural networks for low-level vision. In *CVPR*, 2018. 2

[26] Qingnan Fan, Dongdong Chen, Lu Yuan, Gang Hua, Nenghai Yu, and Baoquan Chen. Decouple learning for parameterized image operators. In *ECCV*, 2018. 2

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

[28] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE TPAMI*, 2019. 2

[29] Hong Wang, Zongsheng Yue, Qi Xie, Qian Zhao, Yefeng Zheng, and Deyu Meng. From rain generation to rain removal. In *CVPR*, 2021. 2

[30] Yuntong Ye, Yi Chang, Hanyu Zhou, and Luxin Yan. Closing the loop: Joint rain generation and removal via disentangled image translation. In *CVPR*, 2021. 2, 6

[31] Hao-Hsiang Yang, Chao-Han Huck Yang, and Yu-Chiang Frank Wang. Wavelet channel attention module with a fusion network for single image deraining. In *ICIP*, 2020. 2

[32] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE TPAMI*, 2010. 2

[33] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *CVPR*, 2020. 2, 3, 6

[34] Qili Deng, Ziling Huang, Chung-Chi Tsai, and Chia-Wen Lin. Hardgan: A haze-aware representation distillation gan for single image dehazing. In *ECCV*, 2020. 2

[35] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *CVPR*, 2020. 2

[36] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, 2020. 2, 6

[37] Xianhui Zheng, Yinghao Liao, Wei Guo, Xueyang Fu, and Xinghao Ding. Single-image-based rain and snow removal using multi-guided filter. In *ICONIP*, 2013. 2

[38] Soo-Chang Pei, Yu-Tai Tsai, and Chen-Yu Lee. Removing rain and snow in a single image using saturation and visibility features. In *ICMEW*, 2014. 2

[39] Yinglong Wang, Shuaicheng Liu, Chen Chen, and Bing Zeng. A hierarchical approach for rain or snow removing in a single color image. *IEEE TIP*, 2017. 2

[40] Shujian Yu, Yixiao Zhao, Yi Mou, Jinghui Wu, Lu Han, Xiaopeng Yang, and Baojun Zhao. Content-adaptive rain and snow removal algorithms for single image. In *ISNN*, 2014. 2

[41] Zhi Li, Juan Zhang, Zhijun Fang, Bo Huang, Xiaoyan Jiang, Yongbin Gao, and Jenq-Neng Hwang. Single image snow removal via composition generative adversarial networks. *IEEE Access*, 7:25016–25025, 2019. 2

[42] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE TIP*, 2021. 2

[43] Fei Yang, Jialu Zhang, and Qian Zhang. Multi-scale capsule generative adversarial network for snow removal. *IET Computer Vision*, 2021. 2

[44] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, 2021. 2, 6

[45] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018. 2

[46] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *CVPR*, 2021. 2

[47] Da-Wei Jaw, Shih-Chia Huang, and Sy-Yen Kuo. Desnowgan: An efficient single image snow removal framework using cross-resolution lateral connection and gans. *IEEE TCSVT*, 2020. 2, 6

[48] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, et al. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[49] Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*, 2018. 3

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

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

[52] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 3

[53] Huafeng Li, Shuanglin Yan, Zhengtao Yu, and Dapeng Tao. Attribute-identity embedding and self-supervised learning for scalable person re-identification. *IEEE TCVST*, 2019. 3

[54] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[55] Peng Xu, Zeyu Song, Qiyue Yin, Yi-Zhe Song, and Liang Wang. Deep self-supervised representation learning for free-hand sketch. *IEEE TCSVT*, 2020. 3

[56] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4

[58] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 4

[59] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 2018. 5

[60] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[61] Jiangxin Dong and Jinshan Pan. Physics-based feature dehazing networks. In *ECCV*. Springer, 2020. 6

[62] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019. 6

[63] 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*, 2017. 6

[64] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *Proc. AAAI Conf. Artif. Intell.*, pages 1–9, 2021. 6