

Video Adverse-Weather-Component Suppression Network via Weather Messenger and Adversarial Backpropagation

Yijun Yang^{1,2}, Angelica I. Aviles-Rivero³, Huazhu Fu⁴, Ye Liu⁵, Weiming Wang⁶, Lei Zhu^{1,2†}

¹The Hong Kong University of Science and Technology (Guangzhou)

²The Hong Kong University of Science and Technology ³University of Cambridge

⁴Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

⁵Tianjin University ⁶Hong Kong Metropolitan University

Abstract

Although convolutional neural networks (CNNs) have been proposed to remove adverse weather conditions in single images using a single set of pre-trained weights, they fail to restore weather videos due to the absence of temporal information. Furthermore, existing methods for removing adverse weather conditions (e.g., rain, fog, and snow) from videos can only handle one type of adverse weather. In this work, we propose the first framework for restoring videos from all adverse weather conditions by developing a video adverse-weather-component suppression network (ViWS-Net). To achieve this, we first devise a weather-agnostic video transformer encoder with multiple transformer stages. Moreover, we design a long short-term temporal modeling mechanism for weather messenger to early fuse input adjacent video frames and learn weather-specific information. We further introduce a weather discriminator with gradient reversion, to maintain the weather-invariant common information and suppress the weather-specific information in pixel features, by adversarially predicting weather types. Finally, we develop a messenger-driven video transformer decoder to retrieve the residual weather-specific feature, which is spatiotemporally aggregated with hierarchical pixel features and refined to predict the clean target frame of input videos. Experimental results, on benchmark datasets and real-world weather videos, demonstrate that our ViWS-Net outperforms current state-of-the-art methods in terms of restoring videos degraded by any weather condition.

1. Introduction

Adverse weather conditions (including rain, fog and snow) often degrade the performance of outdoor vision

systems, such as autonomous driving and traffic surveillance, by reducing environment visibility and corrupting image/video content. Removing these adverse weather effects is challenging yet a promising task. While many video dehazing/deraining/desnowing methods have been proposed, they mainly address one type of weather degradation. As they require multiple models and sets of weights for all adverse weather conditions, resulting in expensive memory and computational costs, they are unsuitable for real-time systems. Additionally, the system would have to switch between a series of weather removal algorithms, making the pipeline more complicated and less practical for real-time systems.

Recently, Li *et al.* [18] proposed an All-in-One bad weather removal network that can remove any weather condition from an image, making it the first algorithm to provide a generic solution for adverse weather removal. Following this problem setting, several single-image multi-adverse-weather removal methods [8, 38] have been developed to remove the degradation effects by one model instance of a single encoder and single decoder. While significant progress has been witnessed for the single-image multi-adverse-weather removal task, we believe that video-level algorithms can achieve better results by utilizing the temporal redundancy from neighboring frames to reduce the inherent ill-posedness in restoration tasks.

Therefore, a generic framework that can transform an image-level algorithm into its video-level counterpart is highly valuable. However, two bottlenecks need to be addressed: 1) how to effectively maintain the temporal coherence of background details across video frames, and 2) how to prevent the perturbation of multiple kinds of weather across video frames.

To tackle the aforementioned bottlenecks, we present the **Video Adverse-Weather-Component Suppression Network (ViWS-Net)**, the first video-level algorithm that can remove all adverse weather conditions with only one set of

[†]Lei Zhu (leizhu@ust.hk) is the corresponding author.

pre-trained weights. Specifically, we introduce Temporally-active Weather Messenger tokens to learn weather-specific information across video frames and retrieve them in our messenger-driven video transformer decoder. We also design a Long Short-term Temporal Modeling mechanism for weather messenger tokens to provide early fusion among frames, and support recovery with temporal dependences of different time spans. To impede the negative effects of multiple adverse weather conditions on background recovery, we develop a Weather-Suppression Adversarial Learning by introducing a weather discriminator. Adversarial backpropagation is adopted, between the video transformer encoder and the discriminator, by gradient reversion to maintain the common background information and simultaneously suppress the weather-specific information in hierarchical pixel features. Since there has been no public dataset for video desnowing, we synthesize the first video-level snow dataset, named KITTI-snow, which is based on KITTI [22]. We conduct extensive experiments on video deraining, dehazing, and desnowing benchmark datasets, including RainMotion [39], REVIDE [49], and KITTI-snow, as well as several real-world weather videos, to validate the effectiveness and generalization of our framework for video multiple adverse weather removal. Our contributions can be summarized as follows:

- We propose a novel unified framework, ViWS-Net, that addresses the problem of recovering video frames from multiple types of adverse weather degradation with a single set of pre-trained weights.
- We introduce temporally-active weather messenger tokens that provide early temporal fusion and help retrieving the residual weather-specific information for consistent removal of weather corruptions.
- We design a weather-suppression adversarial learning approach that maintains weather-invariant background information and suppresses weather-specific information, thereby preventing recovery from the perturbation of various weather types.
- To evaluate our framework under multiple adverse weather conditions, we synthesize a video-level snow dataset KITTI-snow. Our extensive experiments on three benchmark datasets and real-world videos demonstrate the effectiveness and generalization ability of ViWS-Net. Our code is publicly available at <https://github.com/scott-yjyang/ViWS-Net>.

2. Related Work

Video Single-Weather Removal. We briefly introduce different video single-weather removal methods. For video deraining, Garg and Nayar first modeled the video rain and developed a rain detector based on the photometric appearance

of rain streak [12, 13]. Inspired by these seminal works, many subsequent methods focusing on handcrafted intrinsic priors [1–4, 25, 34, 50] have been proposed in the past decades. Recently, deep neural networks have also been employed along this research line [5, 17, 23, 39, 42–45]. Yang *et al.* [43] built a two-stage recurrent network that utilizes dual-level regularizations toward video deraining. Wang *et al.* [39] devised a new video rain model that accounts for rain streak motions, resulting in more accurate modeling of the rain streak layers in videos. For video dehazing, various methods [15, 20, 47] are introduced to generate more accurate dehazed results. For example, with the development of deep learning, Ren *et al.* [33] proposed a synthetic video dehazing dataset and developed a deep learning solution to accumulate information across frames for transmission estimation. To break the limit of poor performance in real-world hazy scenes, Zhang *et al.* [49] developed a video acquisition system that enabled them to capture hazy videos and their corresponding haze-free counterparts from real-world settings. Based on [49], Liu *et al.* [27] proposed a phase-based memory network that integrates color and phase information from the current frame with that of past consecutive frames. For snow removal, while most existing learning-based methods [6, 7, 48] focused on single-image desnowing, no work explored the better solution for video desnowing using temporal information. We propose a novel approach to address the challenge of removing adverse weather effects in videos. Unlike previous methods, we adopt a unified single-encoder single-decoder network that can handle various types of adverse weather conditions using a single model instance.

Single-image Multi-Adverse-Weather Removal. Most recently, a body of researchers has investigated single-image multiple adverse weather removal tasks by one model instance. Li *et al.* [18] developed a single network-based method All-in-One with multiple task-specific encoders and a generic decoder based on Neural Architecture Search (NAS) architecture. It backpropagates the loss only to the respective feature encoder based on the degradation types. TransWeather [38] proposed a transformer-based end-to-end network with only a single encoder and a decoder. It introduced an intra-patch transformer block into the transformer encoder for smaller weather removal. It also utilized a transformer decoder with weather type embeddings learned from scratch to adapt to different weather types. Chen *et al.* [8] proposed a two-stage knowledge distillation mechanism to transfer weather-specific knowledge from multiple well-trained teachers on diverse weather types to one student model. Our study draws attention to multi-adverse-weather removal issue in videos. However, all the above methods failed to capture complementary information from temporal space. Although we can generalize them to remove adverse weather removal in a frame-by-frame

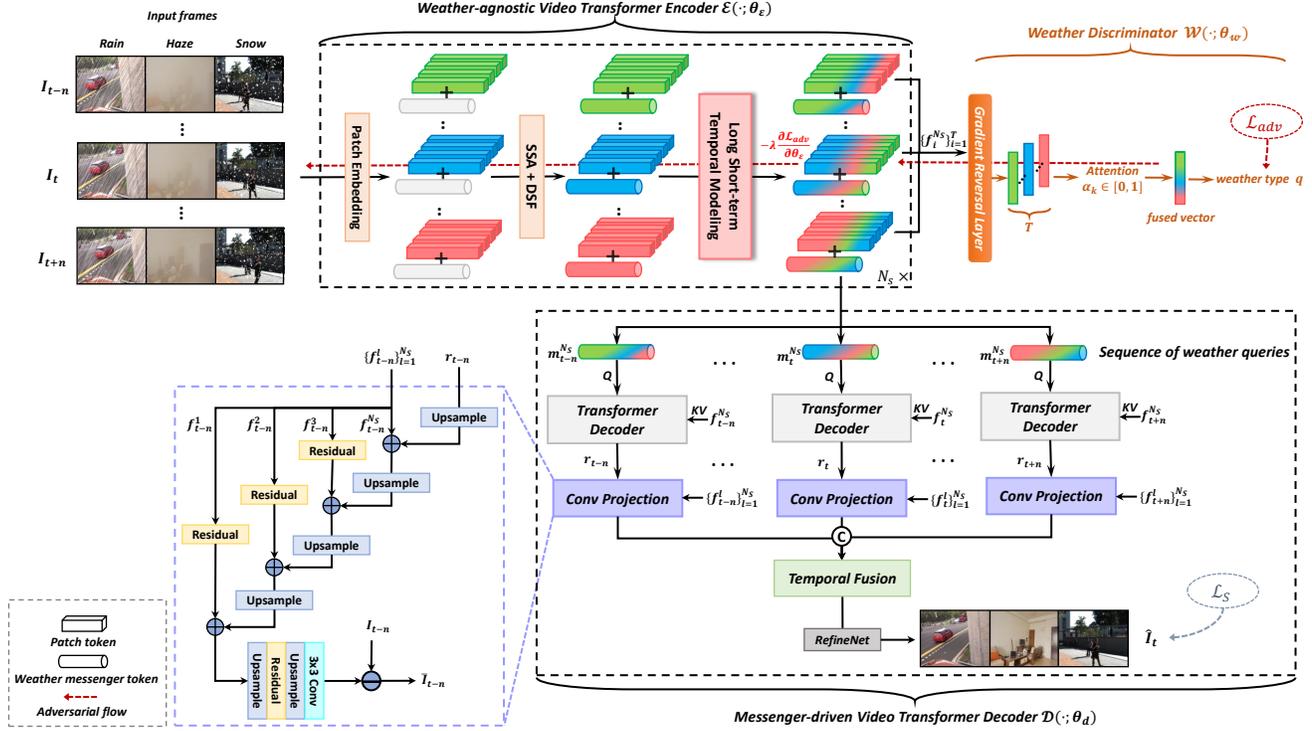


Figure 1. **Overview of our ViWS-Net framework for Video Multiple Adverse Weather Removal.** Given a sequence of video frames, we divide the frames into patch tokens and concatenate them with the corresponding weather messenger token as inputs. The weather messengers temporally collect weather-specific information while the weather-agnostic video transformer encoder performs feature extraction and generates hierarchical pixel features. Simultaneously, a weather discriminator is adversarially learned by the gradient reversal layer to maintain the weather-invariant information and suppress the weather-specific counterpart. For each frame, the messenger-driven video transformer decoder leverages the last pixel feature $f_t^{N_s}$ as key and value, the well-learned weather messenger token $m_t^{N_s}$ as queries to retrieve the weather-specific feature r . Finally, the weather-specific feature r is aggregated together with hierarchical pixel features $\{f^l\}_{l=1}^{N_s}$ across both spatial and temporal axis followed by a refinement network to obtain the final clean target frame \hat{I}_t .

manner, temporal information among video frames enables our method to work better than those image-level ones.

Adversarial Learning. Deep learning has gained popularity in recent years due to its ability to learn non-linear features, making it easier to learn invariant features for multi-task tasks. Adversarial learning, inspired by generative adversarial networks [14], has been employed in natural language processing to learn a common feature representation for multi-task learning, as demonstrated in [24, 28, 36]. These adversarial multi-task models consist of three networks: a feature encoder network, a decoder network, and a domain network. The decoder network minimizes the training loss for all tasks based on the feature encoder network, while the domain network distinguishes the task to which a given data instance belongs. Such learning paradigm has also been used to tackle the domain shift problem [11, 19, 30, 35, 37] to learn domain-invariant information. Inspired by those works, we further explore the common feature representation of multiple adverse weather in videos by adversarial learning paradigm.

3. Method

In this work, our goal is to devise the first video-level unified model to remove multiple types of adverse weather in frames with one set of model parameters. We follow an end-to-end formulation of adverse weather removal as:

$$\hat{I}_t = \mathcal{D}(\mathcal{E}(\mathbf{V}_i^q)), \quad (1)$$

$$\mathbf{V} = \{I_{t-n}, \dots, I_{t-1}, I_t, I_{t+1}, \dots, I_{t+n}\},$$

where \mathbf{V}_i^q is the i -th video clip with $T = 2n + 1$ frames degraded by q -th weather type, \hat{I}_t is the recovered target frame. Different from standard image-level method All-in-One [18], our ViWS-Net tackles multiple adverse weather problem more efficiently by one video transformer encoder $\mathcal{E}(\cdot)$ and one video transformer decoder $\mathcal{D}(\cdot)$. Next, we elaborate our solution for Video Multiple Adverse Weather Removal task.

3.1. Overall Architecture

The overall architecture of our ViWS-Net is displayed in Figure 1, which consists of a weather-agnostic video trans-

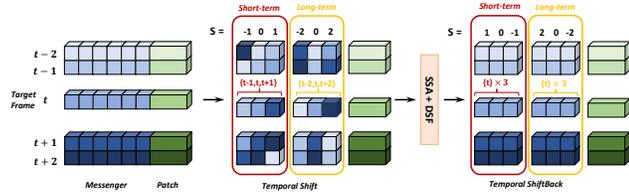


Figure 2. **An illustration of our Long Short-term Temporal Modeling mechanism.** This mechanism is repeatedly applied at each stage of the transformer encoder.

former encoder, a messenger-driven video transformer decoder, and a weather discriminator. Without loss of generality, we build ViWS-Net based on the Shunted transformer [32] consisting of shunted self-attention (SSA) and detail-specific feedforward layer (DSF). SSA extends spatial reduction attention in PVT [40] to unify multi-scale feature extractions within one self-attention layer through multi-scale token aggregation. DSF enhances local details by inserting a depth-wise convolution layer between the two fully connected layers in the feed-forward layer.

Given a sequence of video clip with $T = 2n + 1$ frames $\{I_{t-n}, \dots, I_{t-1}, I_t, I_{t+1}, \dots, I_{t+n}\}$ degraded by q -th adverse weather, our transformer encoder performs feature extraction and generates hierarchical pixel features while weather messenger tokens conduct long short-term temporal modeling for the early fusion in the temporal axis. The weather discriminator with a gradient reversal layer is adversarially learned by predicting the weather type of video clips to maintain the weather-invariant background information and suppress the weather-specific information in the pixel features. The messenger-driven video transformer decoder initializes weather type queries with temporally-active weather messenger well-learned during encoding to retrieve the residual weather-specific information from the suppressed pixel feature. Finally, the hierarchical pixel features and weather-specific feature are spatiotemporally integrated and refined to reconstruct the clean target frame. Empirically, we set $n = 2$ to achieve a good trade-off between performance and computational cost.

3.2. Temporally-Active Weather Messenger

Previous single-image multi-adverse-weather removal work [38] adopted a fixed number of learnable embeddings to query weather-specific features from pixel features in the transformer decoder, termed as weather type queries. However, hindered by random initialization, they are hard to tell the robust weather-specific information during decoding. Furthermore, these query embeddings are independently learned across frames, resulting in the absence of temporal information in the video scenario. To address these limitations, we introduce weather messenger in the video transformer encoder, and the well-learned weather

messengers are adopted as the weather type queries. Specifically, a group of learnable embeddings with size of $M \times C$ is introduced as weather messenger tokens for each frame, which is denoted as $\{m_i^0\}_{i=1}^T \in \mathbb{R}^{T \times M \times C}$. A video clip with the resolution of $H \times W$ is divided and projected into $T \times \frac{HW}{P^2} \times C$ overlapped patch embeddings frame-by-frame, where P and C denote the patch size and the channel dimension respectively. Then, we concatenate patch embeddings of each frame with the corresponding weather messenger tokens before feeding into the video transformer encoder:

$$\{[f_i^0, m_i^0]\}_{i=1}^T \in \mathbb{R}^{T \times (\frac{HW}{P^2} + M) \times C}. \quad (2)$$

The joint tokens $\{[f_i^0, m_i^0]\}_{i=1}^T$ are taken as inputs for the first stage of the transformer encoder. Our video transformer encoder has $N_s = 4$ stages and each stage consists of several blocks of SSA and DSF. The joint token of the l -th stage is learned as:

$$\{[f_i^l, m_i^l]\}_{i=1}^T = \{DSF^l(SSA^l([f_i^{l-1}, m_i^{l-1}]))\}_{i=1}^T. \quad (3)$$

Our weather messengers are temporally active between blocks of each stage to collect weather-specific information from pixel features. To further explore temporal dependence with different spans for the target frame, we conduct a long short-term temporal modeling mechanism as shown in Figure 2. Weather messenger tokens of one frame are separated into 6 groups and shifted along the temporal axis with different time steps (0-2) and directions (forward or backward) followed by an inverse operation (shiftback). For the target frame I_t , the first 3 groups model short-term dependence by shifting messenger tokens of the neighbor frames $\{I_{t-1}, I_{t+1}\}$ with one time step, while the last 3 groups model long-term dependence by shifting messenger tokens of the neighbor frames $\{I_{t-2}, I_{t+2}\}$ with two time steps. Temporal dependences of different spans endow the recovery of the target frame with the comprehensive reference of weather-specific information from past and future frames.

3.3. Weather-Suppression Adversarial Learning

To construct a weather-agnostic transformer encoder, inspired by domain adaptation [11], we design Weather-Suppression Adversarial Learning to learn a great feature space maintaining weather-invariant background information and suppressing weather-specific information. To this end, we optimize a weather discriminator for classifying the weather types by adversarial backpropagation.

Notably, a gradient reversal layer (GRL) is inserted between the video transformer encoder and weather discriminator. During backpropagation, GRL takes the gradient from the weather discriminator, multiplies it by $-\lambda$ and passes it to the transformer encoder. To predict the weather type of a video clip, we combine information from all frames of one video clip by computing an attention-

weighted average of their vector representations. We apply the gated attention mechanism by using the sigmoid function to provide a learnable non-linearity that increases model flexibility. An attention score α_i is computed on each frame as:

$$\alpha_i = \frac{\exp\{\mathbf{w}_1^T(\tanh(\mathbf{w}_2\mathbf{v}_i^T) \cdot \text{sigm}(\mathbf{w}_3\mathbf{v}_i^T))\}}{\sum_{k=1}^T \exp\{\mathbf{w}_1^T(\tanh(\mathbf{w}_2\mathbf{v}_k^T) \cdot \text{sigm}(\mathbf{w}_3\mathbf{v}_k^T))\}}, \quad (4)$$

where $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ are learnable parameters. This process yields an attention-weighted fused vector representation, which reads:

$$\mathbf{v} = \sum_{i=1}^T \alpha_i \mathbf{v}_i, \quad (5)$$

where \mathbf{v}_i is the vector representation from feature embeddings of the frame i . The weather type is finally obtained from the fused vector by one fully connected layer. While the weather discriminator $\mathcal{W}(\cdot)$ seeks an accurate prediction for weather types, the video transformer encoder strives to generate weather-agnostic pixel features. The adversarial loss can be thus achieved by min-max optimization as:

$$\mathcal{L}_{adv} = \min_{\theta_w} \left(\lambda \max_{\theta_\varepsilon} \left(\sum_{q=1}^Q \sum_{i=1}^{N_q} q \log[\mathcal{W}(\mathcal{E}(\mathbf{V}_i^q))] \right) \right). \quad (6)$$

Our weather-suppression adversarial learning develops from the basic idea that the weather-specific information is suppressed in hierarchical pixel features in the transformer encoder by downplaying the discrimination of weather types. This protects the recovery of the target frame from perturbations by different weather types, and thus concentrates the model on the weather-invariant background information. At the training stage, weather-suppression adversarial learning is applied to empower the video transformer encoder with the characteristic of weather-agnostic. At the inference stage, video frames are only fed into the video transformer encoder and decoder for weather removal.

3.4. Messenger-driven Video Transformer Decoder

Intuitively, while weather-suppression adversarial learning largely impedes the appearance of weather-specific information, the residual still may exist in pixel features when the adversarial loss reaches a saddle point. To localize the perturbation from the residual weather-specific information, we design Messenger-driven Video Transformer Decoder to retrieve such information and recover frames from hierarchical features using temporally-active weather messengers described in Section 3.2. Firstly, we adopt the well-learned weather messengers $\{m_i^{Ns}\}_{i=1}^T$ to query the residual weather-specific information. After long short-term temporal modeling in the transformer encoder, weather messengers are trained to locate more true positives of adverse weather in pixel features referring to rich temporal information, than independently-learned query embeddings in

Table 1. The data statistics of RainMotion, REVIDE and KITTI-snow for our video multiple adverse weather removal. The mixed training set is composed of the training set from the three datasets.

Weather	Dataset	Split	Video Num	Video Length	Video Frame Num
Rain	RainMotion	train	40	50	2000
		test	40	20	800
Haze	REVIDE	train	42	7-34	928
		test	6	20-31	154
Snow	KITTI-snow	train	35	50	1750
		test	15	50	750

[38]. With the pixel feature $\{f_i^{Ns}\}_{i=1}^T$ as key and value, the transformer decoder generates the weather-specific feature $\{r_i\}_{i=1}^T$. Note that the transformer decoder here operates at a single stage but has multiple blocks, which are similar to the stage of the transformer encoder. As illustrated in Figure 1, the weather-specific feature is spatially integrated with hierarchical pixel features in the convolution projection block with pairs of an upsampling layer and a 2D convolution residual layers frame-by-frame. To recover details of the background, we subtract the outputs from the original frames. After that, we concatenate the outputs of frames and feed them into the temporal fusion block consisting of three consecutive 3D convolution layers to achieve temporal integration. Finally, we obtain the clean target frame \hat{I}_t by applying a refinement network, which is a vanilla and much smaller version of our ViWS-Net, onto the initial recovered results with tiny artifacts.

The supervised objective function is composed of a smooth L1 loss and a perceptual loss as follows:

$$\mathcal{L}_S = \mathcal{L}_{smoothL1} + \gamma_1 \mathcal{L}_{perceptual}, \quad \text{with} \quad (7)$$

$$\mathcal{L}_{smoothL1} = \begin{cases} 0.5(\hat{I}_t - B_t)^2, & \text{if } |\hat{I}_t - B_t| < 1 \\ |\hat{I}_t - B_t| - 0.5, & \text{otherwise,} \end{cases} \quad (8)$$

$$\mathcal{L}_{perceptual} = \mathcal{L}_{mse}(VGG_{3,8,15}(\hat{I}_t), VGG_{3,8,15}(B_t)), \quad (9)$$

where \hat{I}_t, B_t denote the prediction and ground truth of the target frame, respectively. The overall objective function is composed of supervised loss and adversarial loss, which can be defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_S + \gamma_2 \mathcal{L}_{adv}, \quad (10)$$

where γ_1 and γ_2 are the balancing hyper-parameters, empirically set as 0.04 and 0.001, respectively.

4. Experiments

In this section, we describe in detail the range of experiments that we conducted to validate our proposed method.

4.1. Datasets

Various video adverse weather datasets are used in our experiments. Table 1 summarizes the information of our video multiple adverse weather datasets. RainMotion [39]

Table 2. **Quantitative evaluation for video multiple adverse weather removal.** For Original Weather, these methods are trained on the weather-specific training set and tested on the weather-specific testing set. For Rain, Haze, and Snow, these methods are trained on a mixed training set and tested on the weather-specific testing set. The average performance is calculated on Rain, Haze, and Snow. PSNR and SSIM are adopted as our evaluation metrics. The top values are denoted in red.

Methods		Type	Source	Datasets									
				Original Weather		Rain		Haze		Snow		Average	
Derain	PReNet [31]	Image	CVPR'19	27.06	0.9077	26.80	0.8814	17.64	0.8030	28.57	0.9401	24.34	0.8748
	SLDNet [44]	Video	CVPR'20	20.31	0.6272	21.24	0.7129	16.21	0.7561	22.01	0.8550	19.82	0.7747
	S2VD [45]	Video	CVPR'21	24.09	0.7944	28.39	0.9006	19.65	0.8607	26.23	0.9190	24.76	0.8934
	RDD-Net [39]	Video	ECCV'22	31.82	0.9423	30.34	0.9300	18.36	0.8432	30.40	0.9560	26.37	0.9097
Dehaze	GDN [26]	Image	ICCV'19	19.69	0.8545	29.96	0.9370	19.01	0.8805	31.02	0.9518	26.66	0.9231
	MSBDN [10]	Image	CVPR'20	22.01	0.8759	26.70	0.9146	22.24	0.9047	27.07	0.9340	25.34	0.9178
	VDHNet [33]	Video	TIP'19	16.64	0.8133	29.87	0.9272	16.85	0.8214	29.53	0.9395	25.42	0.8960
	PM-Net [27]	Video	MM'22	23.83	0.8950	25.79	0.8880	23.57	0.9143	18.71	0.7881	22.69	0.8635
Desnow	DesnowNet [29]	Image	TIP'18	28.30	0.9530	25.19	0.8786	16.43	0.7902	27.56	0.9181	23.06	0.8623
	DDMSNET [48]	Image	TIP'21	32.55	0.9613	29.01	0.9188	19.50	0.8615	32.43	0.9694	26.98	0.9166
	HDCW-Net [7]	Image	ICCV'21	31.77	0.9542	28.10	0.9055	17.36	0.7921	31.05	0.9482	25.50	0.8819
	SMGARN [9]	Image	TCSVT'22	33.24	0.9721	27.78	0.9100	17.85	0.8075	32.34	0.9668	25.99	0.8948
Restoration	MPRNet [46]	Image	CVPR'21	—	—	28.22	0.9165	20.25	0.8934	30.95	0.9482	26.47	0.9194
	EDVR [41]	Video	CVPR'19	—	—	31.10	0.9371	19.67	0.8724	30.27	0.9440	27.01	0.9178
	RVRT [21]	Video	NIPS'22	—	—	30.11	0.9132	21.16	0.8949	26.78	0.8834	26.02	0.8972
	RTA [51]	Video	CVPR'22	—	—	30.12	0.9186	20.75	0.8915	29.79	0.9367	26.89	0.9156
All-in-one [18]		Image	CVPR'20	—	—	26.62	0.8948	20.88	0.9010	30.09	0.9431	25.86	0.9130
UVRNet [16]		Image	TMM'22	—	—	22.31	0.7678	20.82	0.8575	24.71	0.8873	22.61	0.8375
TransWeather [38]		Image	CVPR'22	—	—	26.82	0.9118	22.17	0.9025	28.87	0.9313	25.95	0.9152
TKL [8]		Image	CVPR'22	—	—	26.73	0.8935	22.08	0.9044	31.35	0.9515	26.72	0.9165
Ours		Video	—	—	—	31.52	0.9433	24.51	0.9187	31.49	0.9562	29.17	0.9394

is the latest video deraining dataset synthesized based on NTURain [5]. It has five large rain streak masks, making it more demanding to remove the rain streaks. REVIDE [49] is the first real-world video dehazing dataset with high fidelity real hazy conditions recording indoor scenes. To our best knowledge, there have not been any public video-level snow datasets yet. Thus, we built our own video desnowing dataset named KITTI-snow. The details of KITTI-snow are presented as follows. At the training stage, we merge the training set of the three datasets to learn a unified model. For the testing stage, we evaluate our model on three testing sets, respectively.

KITTI-snow: We create a synthesized outdoor dataset called KITTI-snow that comprises 50 videos with a total of 2500 frames, all featuring snowy conditions. Specifically, we randomly collect two groups of videos from KITTI [22]. The first group consists of 35 videos and is treated as the training set, while the second group includes 15 videos and is treated as the testing set. Given each clean video, we synthesize snowflakes with different properties (i.e. transparency, size and position) according to Photoshop’s snow synthesis tutorial. To better simulate the real-world snow scene, gaussian blurring is applied onto snow particles. To model the temporal consistency, we sample the position, size and blurring degree of snow in different frames of the same video from the same distribution. The spatial resolution of video frames is 1000×300 . Figure 3 presents the example frames of five videos with different distributions in our synthetic dataset.

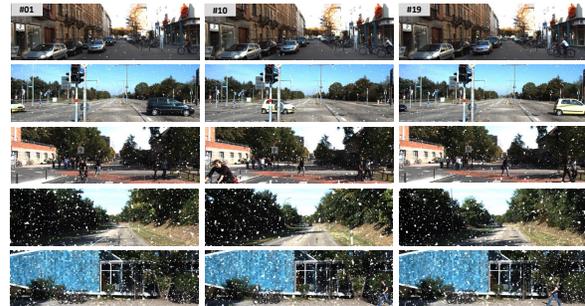


Figure 3. **Example frames of five synthesized videos in KITTI-snow.** The snowflakes in each video are sampled from different distributions.

4.2. Implementation Details

For training details, the proposed framework was trained on two NVIDIA RTX 3090 GPUs and implemented on the Pytorch platform. Our framework is empirically trained for 500 epochs in an end-to-end way and the Adam optimizer is applied. The initial learning rate is set to 2×10^{-4} and decayed by 50% every 100 epochs. We randomly crop the video frames to 224×224 . We empirically set $n = 2$, which means that our network receives 5 frames for each video clip. A batch of 12 video clips evenly composed of three weather types (i.e., rain, haze, snow) is fed into the network for each time.

For method details, the number of weather messenger tokens M for each frame is set to 48. In order to suppress noisy signal from the weather discriminator at the early

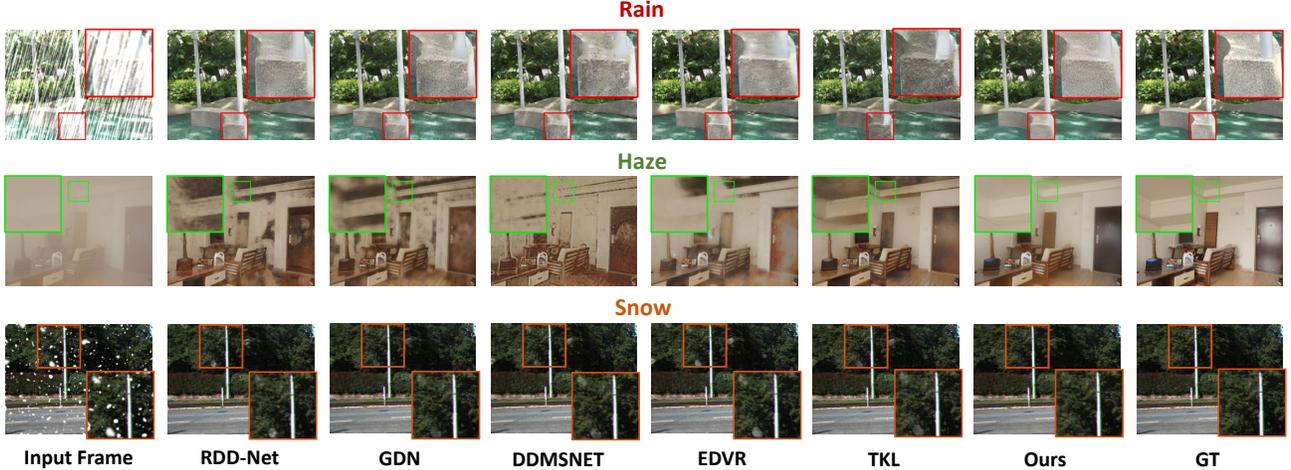


Figure 4. **Qualitative Comparison between adverse weather removal algorithms.** The best algorithms designed for different tasks are selected to present the results on the example frames degraded by rain, haze, snow, respectively. The color box indicates the detailed comparison of weather removal.

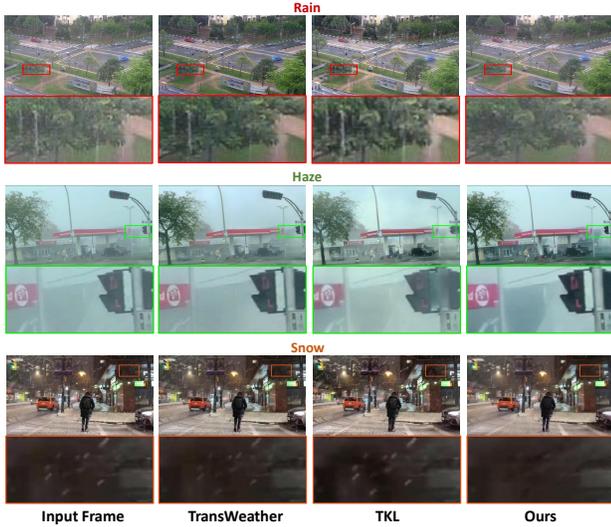


Figure 5. **Visual comparison of different multiple adverse weather removal methods on three real-world video sequences degraded by rain, haze, snow, respectively.** The color boxes display zoom-in views highlighting detailed comparisons of weather removal. Apparently, our network can more effectively remove rain streaks, haze, and snowflakes of input video frames than state-of-the-art methods.

stages of the training procedure, we gradually change the adaptation factor λ from 0 to 1 following the schedule:

$$\lambda = \frac{2}{1 + \exp(-10 \cdot p)} - 1, \quad (11)$$

where p is the current iteration number divided by the total iteration number.

Table 3. Quantitative comparison of computational complexity between the selected models and ViWS-Net. The best values are denoted in bold.

Methods	Parameters (M)	FLOPs (G)	Inference time (s)
TransWeather [38]	24.01	37.68	0.49
TKL [8]	28.71	94.05	0.51
EDVR [41]	20.70	335.27	0.63
ViWS-Net(Ours)	57.82	68.72	0.46

4.3. Quantitative Evaluation

Comparison methods. As shown in Table 2, we compared our proposed method against five kinds of state-of-the-art methods on our mixed dataset. For *derain*, we compared our method with one single-image approach PReNet [31] and three video approaches SLDNet [44], S2VD [45], RDD-Net [39]. For *dehaze*, we compared with two single-image approaches GDN [26], MSBDN [10] and two video approaches VDHNet [33], PM-Net [27]. For *desnow*, we compared with four single-image methods including DesnowNet [29], DDMSNET [48], HDCW-Net [7], SMGARN [9]. For *restoration*, we compared ours with one single-image method MPRNet [46] and three video methods EDVR [41], RVRT [21], RTA [51]. For *multi-adverse-weather removal*, we compared ours with the latest four single-image methods All-in-one [18], UVRNet [16], TransWeather [38], TKL [8].

Analysis on multi-adverse-weather removal. For quantitative evaluation of the restored results, we apply the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) as the metrics. For the single-weather removal models (*derain*, *dehaze*, *desnow*), two types of results are reported: **(i)** the model trained on their original weather (i.e., single weather training set) and **(ii)** the model trained

Table 4. Ablation study of each critical module in the proposed framework on three weather types. The top values are marked in bold font. “WS. Adv.” denote weather-suppression adversarial learning.

Combination	Module			Datasets							
	WeatherMessenger	VideoDecoder	WS. Adv.	Rain		Haze		Snow		Average	
M1	-	-	-	26.92	0.9273	22.77	0.9052	29.94	0.9462	26.54	0.9262
M2	✓	-	-	30.03	0.9327	23.92	0.9149	30.54	0.9520	28.16	0.9332
M3	-	✓	-	29.33	0.9365	22.84	0.9085	30.89	0.9554	27.69	0.9335
M4	-	-	✓	29.70	0.9316	23.87	0.9152	30.82	0.9521	28.13	0.9330
M5	✓	✓	-	31.00	0.9419	24.13	0.9164	30.93	0.9552	28.69	0.9378
Ours	✓	✓	✓	31.52	0.9433	24.51	0.9187	31.49	0.9562	29.17	0.9394

Table 5. Ablation study of the proposed messenger-driven video transformer decoder. The top values are denoted in bold.

Module		Average	
TemporalFusion	RefineNet	PSNR	SSIM
-	-	28.37	0.9305
✓	-	28.80	0.9357
✓	✓	29.17	0.9394

on data of all weather types (i.e., the mixed training set). For restoration and multi-adverse-weather removal models, only the results of the model trained on the mixed training set are reported. For a fair comparison, we retrain each compared model implemented by the official codes based on our training dataset and report the best result. One can see that, our method achieves the best average performance when trained on multi-weather types by a considerable margin of 2.16, 0.0216 in PSNR, SSIM, respectively, than the second-best method EDVR [41]. Although our method may not be the best compared to single-weather removal methods when trained on single-weather data, these methods usually go to failure when coming to multiple adverse weather conditions. For example, while the derain method RDD-Net [39] fails to remove the haze degradation, the dehaze method PM-Net [27] and desnow method DDMSNET [48] have poor performance on snow and haze removal, respectively. Also, it can be observed that DDMSNET [48] and SMGARN [9] still achieve promising results for snow removal when trained on multi-weather types by incorporating snow-specialized modules. However, these methods struggle to address other degradations like haze, leading to lower average performance in multi-weather restoration. In contrast to existing methods, our approach can achieve consistent performance across all weather types by relying solely on a unified architecture and a set of pre-trained weights.

Analysis on computational complexity. We evaluate computational complexity (the number of training parameters, FLOPs, inference time) by feeding a 5-frame video clip with a resolution of 224×224 into our model and the representative models. Our ViWS-Net maintains comparable computational complexity to other methods while achiev-

ing the best results on multi-adverse-weather removal.

4.4. Qualitative Evaluation

Results on our datasets. To better illustrate the effectiveness of our ViWS-Net, Figure 4 shows the visual comparison under our rain, haze, and snow scenarios between our method and 5 state-of-the-art methods that are, respectively, the one with the best average performance for each group of methods. Obviously, one can notice that our method can achieve promising results in visual quality in each weather type. For rain and snow scenarios, the results recovered by our method contain less rain streaks and snow particles compared with other methods. For the hazy scenario, our method can remove more residual haze and much better preserve clean background.

Results on real-world degraded videos. To evaluate the universality of our video multiple adverse weather removal network, we collect three real-world degraded videos, i.e., one rainy video from NTURain*, one hazy video and one snowy video from Youtube website, and further compare our network against state-of-the-art multi-adverse-weather removal methods. Figure 5 shows the visual results produced by our network and two selected methods on real-world video frames. Apparently observed from the detailed comparison, our method outperforms other methods in all weather types by effectively removing adverse weather and maintaining background details.

4.5. Ablation Study

Effectiveness of each module in ViWS-Net. We evaluate the effectiveness of each proposed module including temporally-active weather messenger, video transformer decoder, and weather-suppression adversarial learning (WS. Adv.) as shown in Table 4. We report the result tested on the weather-specific testing set and trained on the mixed training set. The baseline M1, which consists of a Shunted Transformer encoder and a convolution projection decoder, achieves the average performance on three adverse weather datasets of 26.54, 0.9262 in PSNR, SSIM, respectively. M2 introduces temporally-active weather messenger tokens in

*<https://github.com/hotndy/SPAC-SupplementaryMaterials/>

the transformer encoder based on M1 and advances the average performance by 1.62, 0.0070 of PSNR, SSIM, respectively, demonstrating the effectiveness of our proposed Long Short-term Temporal Modeling strategy. M3 presents the messenger-driven video transformer decoder (weather type queries are randomly initialized), while M4 brings in the weather-suppression adversarial learning based on M1. Both M3 and M4 boost the average performance by a significant margin. M5 is developed from M2 and M3, where the weather type queries are initialized by the well-learned weather messenger tokens, leading to a better average performance of 28.69, 0.9378 in PSNR, SSIM. Our full model further applies the weather-suppression adversarial learning strategy and gains a critical increase of 0.48, 0.0016 in PSNR, SSIM, respectively, compared with M5.

Effectiveness of video transformer decoder. We further validate the effectiveness of Temporal Fusion module and RefineNet module in our elaborated video transformer decoder as shown in Table 5. Our reported results were obtained by testing our approach on a mixed testing set and training it on a mixed training set. It is worth noting that both of them benefit the average performance.

5. Conclusion

This paper presents ViWS-Net, an innovative method for simultaneously addressing multiple adverse weather conditions in video frames using a unified architecture and a single set of pre-trained weights. Our approach incorporates Weather-Suppression Adversarial Learning to mitigate the adverse effects of different weather conditions, and Weather Messenger to leverage rich temporal information for consistent recovery. We evaluate our proposed method on benchmark datasets and real-world videos, and our experimental results demonstrate that ViWS-Net achieves superior performance compared to state-of-the-art methods. Ablation studies are also conducted to validate the effectiveness of each proposed module.

Acknowledgments

This work was supported by the Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), National Natural Science Foundation of China (Grant No. 61902275), and Hong Kong Metropolitan University Research Grant (No. RD/2021/09).

References

[1] Peter C Barnum, Srinivasa Narasimhan, and Takeo Kanade. Analysis of rain and snow in frequency space. *International journal of computer vision*, 86(2):256–274, 2010. 2

[2] Jérémie Bossu, Nicolas Hautiere, and Jean-Philippe Tarel. Rain or snow detection in image sequences through use of a

histogram of orientation of streaks. *International journal of computer vision*, 93(3):348–367, 2011.

[3] Nathan Brewer and Nianjun Liu. Using the shape characteristics of rain to identify and remove rain from video. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 451–458. Springer, 2008.

[4] Jie Chen and Lap-Pui Chau. A rain pixel recovery algorithm for videos with highly dynamic scenes. *IEEE transactions on image processing*, 23(3):1097–1104, 2013. 2

[5] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6286–6295, 2018. 2, 6

[6] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Cheng-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 *European Conference on Computer Vision*, pages 754–770. Springer, 2020. 2

[7] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021. 2, 6, 7

[8] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17653–17662, 2022. 1, 2, 6, 7

[9] Bodong Cheng, Juncheng Li, Ying Chen, Shuyi Zhang, and Tiejong Zeng. Snow mask guided adaptive residual network for image snow removal. *arXiv preprint arXiv:2207.04754*, 2022. 6, 7, 8

[10] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020. 6, 7

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3, 4

[12] Kshitiz Garg and Shree K Nayar. Detection and removal of rain from videos. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[13] Kshitiz Garg and Shree K Nayar. Photorealistic rendering of rain streaks. *ACM Transactions on Graphics (TOG)*, 25(3):996–1002, 2006. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

- [15] Jin-Hwan Kim, Won-Dong Jang, Yongsup Park, Dong-Hahk Lee, Jae-Young Sim, and Chang-Su Kim. Temporally x real-time video dehazing. In *2012 19th IEEE International Conference on Image Processing*, pages 969–972. IEEE, 2012. 2
- [16] Ashutosh Kulkarni, Prashant W Patil, Subrahmanyam Murala, and Sunil Gupta. Unified multi-weather visibility restoration. *IEEE Transactions on Multimedia*, 2022. 6, 7
- [17] Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, and Deyu Meng. Video rain streak removal by multiscale convolutional sparse coding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6644–6653, 2018. 2
- [18] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020. 1, 2, 3, 6, 7
- [19] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 3
- [20] Zhuwen Li, Ping Tan, Robby T Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4988–4997, 2015. 2
- [21] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhong Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. *arXiv preprint arXiv:2206.02146*, 2022. 6, 7
- [22] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 6
- [23] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3233–3242, 2018. 2
- [24] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*, 2017. 3
- [25] Peng Liu, Jing Xu, Jiafeng Liu, and Xianglong Tang. Pixel based temporal analysis using chromatic property for removing rain from videos. *Comput. Inf. Sci.*, 2(1):53–60, 2009. 2
- [26] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Grid-hazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7314–7323, 2019. 6, 7
- [27] Ye Liu, Liang Wan, Huazhu Fu, Jing Qin, and Lei Zhu. Phase-based memory network for video dehazing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5427–5435, 2022. 2, 6, 7, 8
- [28] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wascell. Multi-task adversarial network for disentangled feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3743–3751, 2018. 3
- [29] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 6, 7
- [30] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Bakhtashmotlagh, and Sridha Sridharan. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020. 3
- [31] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 6, 7
- [32] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2022. 4
- [33] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE transactions on image processing*, 28(4):1895–1908, 2018. 2, 6, 7
- [34] Varun Santhaseelan and Vijayan K Asari. Utilizing local phase information to remove rain from video. *International Journal of Computer Vision*, 112(1):71–89, 2015. 2
- [35] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 3
- [36] Yusuke Shinohara. Adversarial multi-task learning of deep neural networks for robust speech recognition. In *Inter-speech*, pages 2369–2372. San Francisco, CA, USA, 2016. 3
- [37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 3
- [38] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 1, 2, 4, 5, 6, 7
- [39] Shuai Wang, Lei Zhu, Huazhu Fu, Jing Qin, Carola-Bibiane Schönlieb, Wei Feng, and Song Wang. Rethinking video rain streak removal: A new synthesis model and a deraining network with video rain prior. In *European Conference on Computer Vision*, pages 565–582. Springer, 2022. 2, 5, 6, 7, 8
- [40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 4
- [41] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and

- Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6, 7, 8
- [42] Wending Yan, Robby T Tan, Wenhan Yang, and Dengxin Dai. Self-aligned video deraining with transmission-depth consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11966–11976, 2021. 2
- [43] Wenhan Yang, Jiaying Liu, and Jiashi Feng. Frame-consistent recurrent video deraining with dual-level flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1670, 2019. 2
- [44] Wenhan Yang, Robby T Tan, Shiqi Wang, and Jiaying Liu. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1720–1729, 2020. 6, 7
- [45] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng. Semi-supervised video deraining with dynamical rain generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–652, 2021. 2, 6, 7
- [46] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 6, 7
- [47] Jiawan Zhang, Liang Li, Yi Zhang, Guoqiang Yang, Xiaochun Cao, and Jizhou Sun. Video dehazing with spatial and temporal coherence. *The Visual Computer*, 27(6):749–757, 2011. 2
- [48] 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 Transactions on Image Processing*, 30:7419–7431, 2021. 2, 6, 7, 8
- [49] Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9239–9248, 2021. 2, 6
- [50] Xiaopeng Zhang, Hao Li, Yingyi Qi, Wee Kheng Leow, and Teck Khim Ng. Rain removal in video by combining temporal and chromatic properties. In *2006 IEEE international conference on multimedia and expo*, pages 461–464. IEEE, 2006. 2
- [51] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting temporal alignment for video restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6053–6062, 2022. 6, 7