Video Dehazing via a Multi-Range Temporal Alignment Network with Physical Prior

Jiaqi Xu [1, 2, ⋆], Xiaowei Hu [2, ✉], Lei Zhu [3, 4, ✉], Qi Dou [1], Jifeng Dai [5, 2], Yu Qiao [2], Pheng-Ann Heng [1]
[1] The Chinese University of Hong Kong [2] Shanghai Artificial Intelligence Laboratory
[3] The Hong Kong University of Science and Technology (Guangzhou)
[4] The Hong Kong University of Science and Technology [5] Tsinghua University

Abstract

Video dehazing aims to recover haze-free frames with high visibility and contrast. This paper presents a novel framework to effectively explore the physical haze priors and aggregate temporal information. Specifically, we design a memory-based physical prior guidance module to encode the prior-related features into long-range memory. Besides, we formulate a multi-range scene radiance recovery module to capture space-time dependencies in multiple space-time ranges, which helps to effectively aggregate temporal information from adjacent frames. Moreover, we construct the first large-scale outdoor video dehazing benchmark dataset, which contains videos in various real-world scenarios. Experimental results on both synthetic and real conditions show the superiority of our proposed method.

1. Introduction

Haze largely degrades the visibility and contrast of the outdoor scenes, which adversely affects the performance of downstream vision tasks, such as the detection and segmentation in autonomous driving and surveillance. According to the atmospheric scattering model [18, 38], the formation of a hazy image is described as:

$$I(x) = J(x)t(x) + A(1 - t(x)) \ , \qquad (1)$$

where I, J, A, t denote the observed hazy image, scene radiance, atmospheric light, and transmission, respectively, and x is the pixel index. The transmission $t = e^{-\beta d(x)}$ describes the scene radiance attenuation caused by the light scattering, where β is the scattering coefficient of the atmosphere, and d denotes the scene depth.

Video dehazing benefits from temporal clues, such as highly correlated haze thickness and lighting conditions, as well as the moving foreground objects and backgrounds. Early deep learning-based video dehazing methods leverage temporal information by simply concatenating input frames or feature maps [27, 46]. Recently, GC-IDN [59] proposes to use cost volume and confidence to align and aggregate temporal information. However, existing video dehazing methods suffer from several limitations. First, these approaches either obtain haze-free frames from the physical model-based component estimation [27, 46] or ignore the explicit physical prior embedded in the haze imaging model [59]. The former suffers from inaccurate intermediate prediction, thus leading to error accumulation in the final results, while the latter overlooks the physical prior information, which plays an important role in haze estimation and scene recovery. Second, these methods aggregate temporal information by using input/feature stacking or frame-to-frame alignment in a local sliding window, which is hard to obtain global and long-range temporal information.

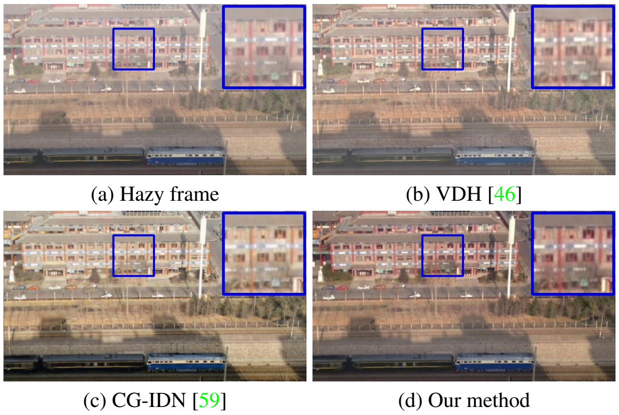

(a) Hazy frame
(b) VDH [46]
(c) CG-IDN [59]
(d) Our method

Figure 1. Visual comparison on a real-world hazy video. Our method trained on our outdoor dehazing dataset clearly removes haze without color distortion.

In this work, we present a novel video dehazing framework via a Multi-range temporal Alignment network with Physical prior (MAP-Net) to address the aforementioned issues. First, we design a memory-based physical prior guid-

⋆ : This work was done during Jiaqi Xu's internship at Shanghai Artificial Intelligence Laboratory.
✉ : Corresponding authors (huxiaowei@pjlab.org.cn; leizhu@ust.hk).

ance module, which aims to inject the physical prior to help the scene radiance recovery. Specifically, we perform feature disentanglement according to the physical model with two decoders, where one estimates the transmission and atmospheric light, and the other recovers scene radiance. The feature extracted from the first decoder is leveraged as the physical haze prior, which is integrated into the second decoder for scene radiance recovery. To infer the global physical prior in a long-range video, we design a physical prior token memory that effectively encodes prior-related features into compact tokens for efficient memory reading.

Second, we introduce a multi-range scene radiance recovery module to capture space-time dependencies in multiple space-time ranges. This module first splits the adjacent frames into multiple ranges, then aligns and aggregates the corresponding recurrent range features, and finally recovers the scene radiance. Unlike CG-IDN [59], which aligns the adjacent features frame-by-frame, we align the features of adjacent frames into multiple sets with different ranges, which helps to explore the temporal haze clues in various time intervals. We further design a space-time deformable attention to warp the features of multiple ranges to the target frame, followed by a guided multi-range complementary information aggregation. Also, we use an unsupervised flow loss to encourage the network to focus on the aligned areas and train the whole network in an end-to-end manner.

In addition, the existing learning-based video dehazing methods are mainly trained and evaluated on indoor datasets [27, 46, 59], which suffer from performance degradation in real-world outdoor scenarios. Thus, we construct an outdoor video dehazing benchmark dataset, HazeWorld, which has three main properties. First, it is a large-scale synthetic dataset with 3,588 training videos and 1,496 testing videos. Second, we collect videos from diverse outdoor scenarios, *e.g*., autonomous driving and life scenes. Third, the dataset has various downstream tasks for evaluation, such as segmentation and detection. Various experiments on both synthetic and real datasets demonstrate the effectiveness of our approach, which clearly outperforms the existing image and video dehazing methods; see Fig. 1. The code and dataset are publicly available at https://github.com/jiaqixuac/MAP-Net.

Our main contributions are summarized as follows:

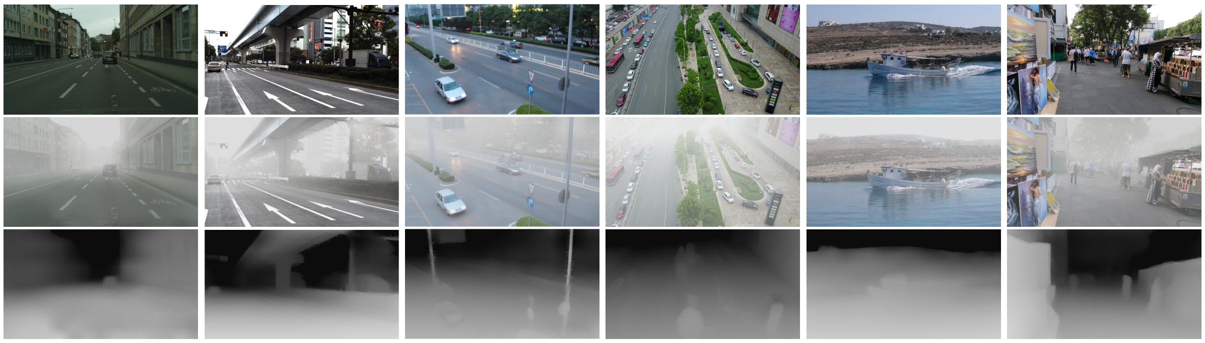
- We present a novel framework, MAP-Net, for video dehazing. A memory-based physical prior guidance module is designed to enhance the scene radiance recovery, which encodes haze-prior-related features into long-range memory.
- We introduce a recurrent multi-range scene radiance recovery module with the space-time deformable attention and the guided multi-range aggregation, which effectively captures long-range temporal haze and scene clues from the adjacent frames.
- We construct a large-scale outdoor video dehazing dataset with diverse real-world scenarios and labels for downstream task evaluation.
- Extensive experiments on both synthetic and real conditions demonstrate our superior performance against the recent state-of-the-art methods.

2. Related Work

Image dehazing. Single-image dehazing has been widely studied in computer vision and computer graphics. Early methods rely on the atmospheric scattering model and physical priors [2, 18]. Later, deep learning-based methods show superior performance by leveraging large numbers of clear/hazy images [1, 28]. These methods either predict the components of the haze physical model [4, 26, 45, 58] or directly restore the haze-free images in an image-to-image translation manner [29, 43] using convolutional neural networks (CNNs). Recent works propose more advanced network and module designs to improve the dehazing performance [9, 12, 13, 17, 33, 35, 42, 49]. However, applying image dehazing methods to videos leads to discontinuous results since the temporal information is simply ignored.

Video dehazing. Video dehazing methods leverage temporal information from the adjacent frames to enhance the restoration quality. Early methods mainly focus on post-processing to generate temporally consistent results by refining transmission maps and suppressing artifacts [8, 22] or joint estimating depths from videos [30]. Li *et al*. [27] present a CNN to optimize dehazing and detection in videos end-to-end. Ren *et al*. [46] use semantic information to regularize the estimated transmission and to improve the video dehazing performance. More recently, Zhang *et al*. [59] collect a real indoor video dehazing dataset (REVIDE) and present a confidence-guided and improved deformable network. Liu *et al*. [34] design a phase-based memory network for video dehazing. Additionally, a neural compression-based method [19] for video restoration shows better performance on REVIDE. However, these methods are mainly trained and evaluated in indoor scenes, and their performance in complex outdoor scenarios is limited.

Video alignment. Alignment aims at obtaining spatial transformation and pixel-wise correspondence from the misaligned frames. Video restoration methods rely on explicit optical flow estimation [44] to align the adjacent images/features [6, 19, 57]. Other methods [51, 53] leverage deformable convolutions [11] to learn the offsets for feature alignment. These methods usually perform the frame-to-frame alignment. More recently, attention [14, 52, 56] with a large receptive field has been used together with the optical flow for feature alignment [5, 31, 32]. Besides, STTN [23] also considers multiple frames but only transforms the input images at one space-time range.



(a) Cityscapes (b) DDAD (c) UA-DETRAC (d) VisDrone (e) DAVIS (f) REDS

Figure 2. Example ground-truth images (the first row), synthetic hazy images (the second row), and transmission maps (the last row) in our HazeWorld dataset.

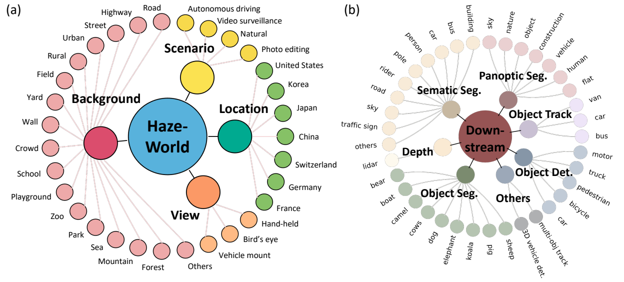


Figure 3. Dataset analysis of our HazeWorld, which contains diverse scenarios and supports various downstream evaluations.

3. HazeWorld Dataset

Since the current video dehazing datasets are mostly collected in indoor scenes, we construct a large-scale synthetic outdoor video dehazing dataset named HazeWorld, with example frames shown in Fig. 2.

Data collection. The original videos of HazeWorld are from six existing datasets, *i.e.*, Cityscapes [10], DDAD [16], UA-DETRAC [54], VisDrone [62], DAVIS [41], and REDS [39], resulting in 1,271 haze-free videos. We use the atmospheric scattering model Eq. (1) to synthesize hazy videos. The robust video depth estimation method [24] is used to obtain temporally consistent depth maps. We follow [3, 47] and choose $\beta \in \{0.005, 0.01, 0.02, 0.03\}$ to generate transmission t, and randomly select $A \in [0.75, 1.0]$ for each video. We split 1,271 haze-free videos into training (897 videos) and testing (374 videos) sets. Overall, we obtain 3,588 and 1,496 hazy synthetic videos with four β of around 240,000 and 86,000 frames in training and testing sets, respectively.

Dataset analysis. As shown in Fig. 3, our dataset contains diverse real-world scenarios, which enables us to assess dehazing performance on various outdoor applications, such as autonomous driving [10, 16], video surveillance [54, 62], and photo editing [39]. Further, the original datasets contain the labels of multiple video and image downstream scene understanding tasks, *e.g.*, video panoptic segmentation [21], object segmentation [41], depth estimation [16], and image semantic segmentation [10]. Thus we can evaluate the effectiveness of dehazing on high-level vision tasks.

4. Methodology

4.1. Overall Framework

Fig. 4 illustrates the overall framework of the proposed MAP-Net, which is a U-Net-like structure that mainly consists of an encoder, a prior decoder, and a scene decoder. A common image backbone, *e.g.*, ConvNeXt [37], is used as the feature encoder, which extracts the multi-scale feature maps. At each scale, features are processed interactively in the prior decoder layer and scene decoder layers. The initial prior feature $\tilde{\mathcal{P}}$ and initial scene feature $\tilde{\mathcal{J}}$ are first fed into a Memory-based Physical prior Guidance (MPG) module (see Sec. 4.2), which aims to obtain the memory-enhanced prior feature $\mathcal{P}$ and the prior-guided scene feature $\mathcal{J}$. Then, $\mathcal{P}$ and $\mathcal{J}$ are fed into a Multi-range Scene radiance Recovery (MSR) module (see Sec. 4.3) , which is to obtain the feature for the haze-free scene by aligning and aggregating recurrent temporal features from the adjacent frames. The prior decoder and scene decoder jointly perform feature disentanglement according to the physical model.

Specifically, the prior decoder learns the prior-related feature by predicting the transmission and atmospheric light, and the scene decoder generates the scene radiance. The intermediate components are obtained using separate prediction heads and reconstructing the hazy input via Eq. (1), which are supervised by a physical model disentanglement loss. Moreover, pixel shuffle layers [48] are used to upsample the features in two decoders and the output from the last scene decoder layer. Lastly, residual prediction is used to produce the final dehazed result.

4.2. Memory-Based Physical Prior Guidance

We design a Memory-based Physical prior Guidance (MPG) module to enhance the scene recovery by encoding haze prior-related features into the long-range memory. Fig. 5 shows the architecture of MPG with three parts.

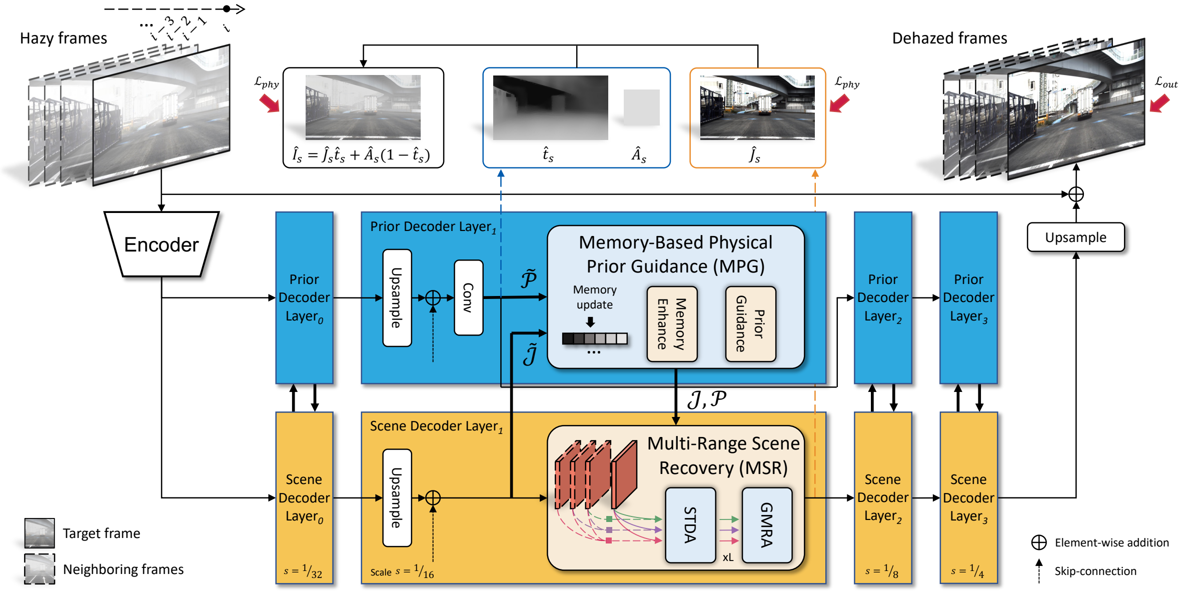


Figure 4. The overall framework of our MAP-Net for video dehazing. MAP-Net is a U-Net-like structure that mainly contains an encoder, a prior decoder, and a scene decoder. Features are processed interactively in the prior decoder and scene decoder, which jointly perform feature disentanglement. The former produces the prior guidance with a memory, and the latter recovers the scene recurrently.

Physical prior compression. The initial prior feature $\tilde{\mathcal{P}} \in \mathbb{R}^{H\times W\times C}$ is implicitly learned using several convolution layers on the upsampled features to predict the transmission map and atmospheric light. H, W, and C denote the feature height, width, and channel size. Since we aim to save the physical priors at different times in the memory, we need to compress the size of each prior to reduce memory space. To achieve this, we first perform a discretization operation on the prior feature by using categorical classification [15, 20] and then normalize the results via the Softmax function. Specifically, from the initial prior, we generate the transmission distribution map $\mathcal{D} \in \mathbb{R}^{H\times W\times D}$, where D is the number of transmission categories. After that, we perform matrix multiplication between the initial prior $\tilde{\mathcal{P}}$ and the transmission distribution map $\mathcal{D}$ to obtain the value $\mathbf{p} \in \mathbb{R}^{D\times C}$, which is the compressed prior token.

Memory-enhanced prior. After obtaining the compressed prior token $\mathbf{p}$, we formulate a prior token memory by saving multiple prior tokens at different time slots. Then, we obtain the feature vectors K and V with the dimension of $\mathbb{R}^{ND\times C}$, where N denotes the number of prior tokens. Hence, we are able to record the historical haze information in video sequences. To perform the interaction between the current haze information and the history information encoded in prior token memory, we adopt the attention operation to read the memory information:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{c}})V\ , \qquad (2)$$

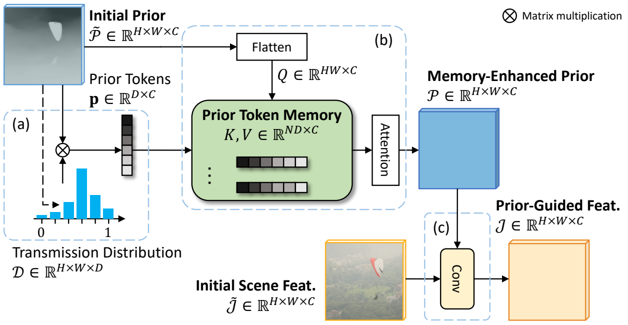

Initial Prior
$\tilde{\mathcal{P}} \in \mathbb{R}^{H\times W\times C}$
Matrix multiplication
Flatten
(b)
$Q \in \mathbb{R}^{HW\times C}$
Prior Tokens
$\mathbf{p} \in \mathbb{R}^{D\times C}$
(a)
Prior Token Memory
$K, V \in \mathbb{R}^{ND\times C}$
Attention
Memory-Enhanced Prior
$\mathcal{P} \in \mathbb{R}^{H\times W\times C}$
0
1
Transmission Distribution
$\mathcal{D} \in \mathbb{R}^{H\times W\times D}$
Initial Scene Feat.
$\tilde{\mathcal{J}} \in \mathbb{R}^{H\times W\times C}$
(c)
Conv
Prior-Guided Feat.
$\mathcal{J} \in \mathbb{R}^{H\times W\times C}$

Figure 5. The illustration of the Memory-based physical Prior Guidance (MPG) module, which has (a) physical prior compression, (b) memory-enhanced prior, and (c) prior feature guidance.

where the query Q is obtained by flatting the initial prior $\tilde{\mathcal{P}}$, and c is the normalization factor, which is the dimension of Q and K. By doing so, we obtain the final memory-enhanced prior $\mathcal{P} = \text{Attention}(Q,K,V)$.

Prior feature guidance. The prior-guided scene feature $\mathcal{J}$ is obtained using several convolution layers, which take the concatenation of the memory-enhanced prior feature $\mathcal{P}$ and the initial scene feature $\tilde{\mathcal{J}}$ as input. Hence, the prior is integrated for scene recovery.

4.3. Multi-Range Scene Radiance Recovery

The Multi-range Scene radiance Recovery module (MSR) aims to capture space-time dependencies in multiple space-time ranges. Fig. 6 shows the detailed structure of our MSR, which aligns the features of adjacent

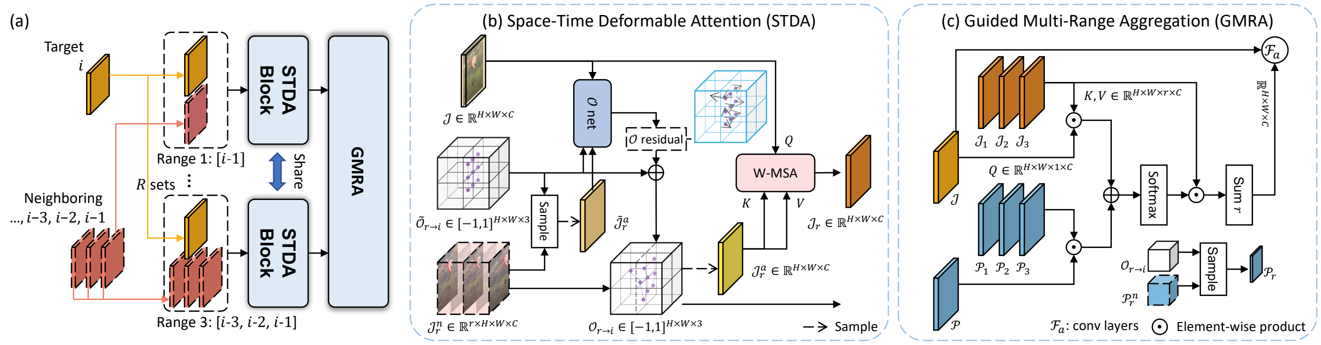

(a)
Target
i
Neighboring
..., i−3, i−2, i−1
STDA Block
STDA Block
Range 1: [i-1]
R sets
Share
Range 3: [i-3, i-2, i-1]
GMRA
(b) Space-Time Deformable Attention (STDA)
$\mathcal{J} \in \mathbb{R}^{H\times W\times C}$
$\mathcal{O}$ net
$\mathcal{O}$ residual
Q
W-MSA
K
V
$\tilde{\mathcal{O}}_{r\to i} \in [-1,1]^{H\times W\times 3}$
Sample
$\tilde{\mathcal{J}}_r^a$
$\mathcal{J}_r \in \mathbb{R}^{H\times W\times C}$
$\mathcal{J}_r^a \in \mathbb{R}^{H\times W\times C}$
$\mathcal{J}_r^n \in \mathbb{R}^{r\times H\times W\times C}$
$\mathcal{O}_{r\to i} \in [-1,1]^{H\times W\times 3}$
Sample
(c) Guided Multi-Range Aggregation (GMRA)
$\mathcal{F}_a$
$K, V \in \mathbb{R}^{H\times W\times r\times C}$
$\mathbb{R}^{H\times W\times C}$
$\mathcal{J}_1$ $\mathcal{J}_2$ $\mathcal{J}_3$
$Q \in \mathbb{R}^{H\times W\times 1\times C}$
$\mathcal{J}$
Softmax
Sum r
$\mathcal{P}_1$ $\mathcal{P}_2$ $\mathcal{P}_3$
$\mathcal{O}_{r\to i}$
Sample
$\mathcal{P}_r$
$\mathcal{P}_r^n$
$\mathcal{P}$
$\mathcal{F}_a$: conv layers
⊙ Element-wise product

Figure 6. The illustration of our Multi-range Scene radiance Recovery (MSR) module. (a) MSR aligns the features of neighboring frames into multiple sets with different ranges. (b) The space-time deformable attention (STDA) block aligns the features of different ranges to the target frame. (c) The guided multi-range aggregation (GMRA) block aggregates the aligned features from multiple sets.

frames $\mathcal{J}_{[i-1,i-2,i-3,\dots]}$ into multiple sets with different ranges, *i.e.*, $\mathcal{J}_{\{[i-1],[i-1,i-2],[i-1,i-2,i-3],\dots\}}$, to explore the temporal haze clues in various time intervals. As shown in Fig. 6 (a), the concatenated features with different ranges are sent to the shared space-time deformable attention block (STDA), which warps the features to the target frame. After that, we formulate a guided multi-range aggregation block (GMRA) to aggregate the aligned features from multiple sets with the guidance of prior features.

4.3.1 Space-Time Deformable Attention

As shown in Fig. 6 (b), the space-time deformable attention (STDA) block aligns the concatenated features of the adjacent frames $\mathcal{J}_r^n = \mathcal{J}_{[i-1,\dots,i-r]} \in \mathbb{R}^{r\times H\times W\times C}$ for each range $r \in \{1,\dots,R\}$ towards the target frame feature $\mathcal{J}$. The output of the STDA block is the range feature $\mathcal{J}_r \in \mathbb{R}^{H\times W\times C}$. Here, we further learn a space-time flow, which is used to capture the correspondence from the previous frames to the current frame. The input space-time flow of the current STDA block is $\tilde{\mathcal{O}}_{r\to i}$, which is gradually refined in this block to produce the output space-time flow $\mathcal{O}_{r\to i}$, following SPy-Net [44].

Specifically, given the concatenated features $\mathcal{J}_r^n$ and a normalized initial space-time flow $\tilde{\mathcal{O}}_{r\to i} \in [-1,1]^{H\times W\times 3}$, we first compute the initial aligned feature map $\tilde{\mathcal{J}}_r^a \in \mathbb{R}^{H\times W\times C}$ as follows:

$$\tilde{\mathcal{J}}_r^a = \mathcal{S}(\mathcal{J}_r^n, \tilde{\mathcal{O}}_{r\to i}) \ , \tag{3}$$

where $\mathcal{S}$ denotes the differentiable space-time sampling operation [23]. Note that the third dimension of $\tilde{\mathcal{O}}_{r\to i} \in [-1,1]^{H\times W\times 3}$ is three, which means the space-time flow capture locations on both spatial domain and time slot. Then, we obtain the refined space-time flow $\mathcal{O}_{r\to i}$ by computing the flow offset residual:

$$\mathcal{O}_{r\to i} = \mathcal{F}_o([\mathcal{J}, \tilde{\mathcal{J}}_r^a, \tilde{\mathcal{O}}_{r\to i}]) + \tilde{\mathcal{O}}_{r\to i} \ , \tag{4}$$

where $\mathcal{F}_o$ is a lightweight offset network composed of convolution layers. Finally, we obtain the aligned feature $\mathcal{J}_r^a$ using Eq. (3) with $\mathcal{O}_{r\to i}$ as input instead.

Cross-attention [52] is used to extract the temporal information from the aligned feature. The feature $\mathcal{J}$ is used as the query $Q = \mathcal{J}U_q$ for the target frame, and the aligned feature $\mathcal{J}_r^a$ is used as the key and value $[K, V] = \mathcal{J}_r^a U_{kv}$, where $U_q \in \mathbb{R}^{C\times C}$, $U_{kv} \in \mathbb{R}^{C\times 2C}$ are learnable projection matrices. Finally, the range feature $\mathcal{J}_r$ is computed as:

$$\mathcal{J}_r = \text{W-MSA}(Q, K, V) \ , \tag{5}$$

where the window multi-head self attention (W-MSA) [36] is leveraged for efficient computation; see Fig. 6 (b). Note that we further adopt the feed-forward network (FFN) to process $\mathcal{J}_r$ after the W-MSA, following [36].

4.3.2 Guided Multi-Range Aggregation

The aligned features of different ranges contain their specific space-time haze clues, and GMRA aggregates multi-range features under the guidance of prior features. Fig. 6 (c) shows the detailed structure, where we compute the aggregation weights from two perspectives, *i.e.*, scene radiance and physical prior. First, the concatenated range features $\{\mathcal{J}_r\}_{r=1}^R$ are considered as the key and value, which are multiplied by the target frame query feature $\mathcal{J}$. For each location, attention weights are computed along the range (r) dimension. Then, the prior guidance is leveraged by computing its derived attention weights in the same way. In specific, we consider the prior feature $\mathcal{P}$ as query and obtain the aligned prior features $\{\mathcal{P}_r\}_{r=1}^R$ as the attention key using Eq. (3) by taking the prior features $\mathcal{P}_r^n = \mathcal{P}_{[i-1,\dots,i-r]}$ and the refined space-time flow $\mathcal{O}_{r\to i}$ as the inputs. The final attention weight is the summation of the weights generated from the scene and prior aspects, followed by a Softmax normalization function for normalization. Finally, multi-range values are aggregated by performing the final attention weight on the range features, which is further summed along the r (range) dimension; see Fig. 6 (c).

Table 1. **Quantitative comparison with state-of-the-art methods on our HazeWorld dataset. Bold** and $\underline{\text{underline}}$ indicate the best and the second-best performance, respectively.

Method	DCP [18]	AOD [26]	GDN [33]	DM2F [12]	FFA [42]	MSBDN [13]	UHD [61]	AECR [55]	Dehamer [17]
PSNR ↑	16.49	15.46	22.80	24.54	22.11	23.70	19.43	22.04	22.92
SSIM ↑	0.8126	0.7997	0.9217	0.9130	0.9007	0.8858	0.7807	0.9067	0.9044
Method	DehazeFormer [49]	EVD [27]	VDH [46]	CG-IDN [59]	FastDVD [50]	EDVR [53]	NCFL [19]	BasicVSR++ [7]	Our method
PSNR ↑	25.44	15.91	17.97	25.25	21.25	22.91	24.33	$\underline{26.06}$	**27.12**
SSIM ↑	$\underline{0.9286}$	0.7968	0.7780	0.9155	0.8678	0.9036	0.9253	0.9207	**0.9349**

Table 2. **Quantitative comparison on the REVIDE dataset [59].** Other baseline results are provided by NCFL [19] paper.

Method	DCP [18]	GDN [33]	MSBDN [13]	FFA [42]	VDH [46]	EDVR [53]	CG-IDN [59]	NCFL [19]	BasicVSR++ [7]	Our method
PSNR ↑	11.03	19.69	22.01	16.65	16.64	21.22	23.21	$\underline{23.63}$	21.68	**24.16**
SSIM ↑	0.7285	0.8545	0.8759	0.8133	0.8133	0.8707	0.8836	$\underline{0.8925}$	0.8726	**0.9043**

4.4. Loss Functions

The overall loss $\mathcal{L}$ is the summation of an output loss $\mathcal{L}_{out}$, a physical model disentanglement loss $\mathcal{L}_{phy}$, and a flow loss $\mathcal{L}_{flow}$:

$$\mathcal{L} = \mathcal{L}_{out} + \lambda_{phy}\mathcal{L}_{phy} + \lambda_{flow}\mathcal{L}_{flow}, \tag{6}$$

where $\lambda_{rec}, \lambda_{flow}$ are the weighting hyper-parameters.

The output loss $\mathcal{L}_{out} = \mathcal{L}_1(\hat{J}, J)$ supervises the final dehazed results $\hat{J}$ with the ground truth J. The physical model disentanglement loss $\mathcal{L}_{phy} = \sum_{s=0}^{3} 2^{s-3}\mathcal{L}_1(\hat{I}_s, I_s) + \mathcal{L}_1(\hat{J}_s, J_s)$ is to make the prior decoder and scene decoder learn the physical model-based components at each scale s in the U-Net, by predicting $\hat{t}_s$, $\hat{A}_s$, and $\hat{J}_s$, and reconstructing input $\hat{I}_s$ using Eq. (1). Moreover, to make the STDA attend to informative regions, we use an unsupervised flow loss to regularize the learned space-time flow in Sec. 4.3.1. Specifically, the flow loss $\mathcal{L}_{flow} = \sum_{s=0}^{3}\sum_{r=1}^{R} 2^{s-3}\mathcal{L}_1(\hat{J}_{sr}^{a}, J_s)$ computes the difference between the warped image $\hat{J}_{sr}^{a}$ and the reference ground truth frame J_s with the scale s for each range r. The warped image is obtained by Eq. (3) with the adjacent ground truth frames and the learned space-time flow as the inputs.

5. Experimental Results

5.1. Settings

Datasets. We evaluate the effectiveness of the proposed MAP-Net on our dataset, *i.e*., HazeWorld, and the widely-used REVIDE dataset [59]. HazeWorld contains 3,588 training videos and 1,496 testing videos. Meanwhile, REVIDE consists of 42 training videos and 6 testing videos.

Evaluation metrics. We utilize PSNR and SSIM to quantitatively evaluate the dehazing performance.

Comparison methods. On HazeWorld, we compare our method against state-of-the-art methods, including ten image dehazing methods (*i.e*., DCP [18], AOD [26], GDN [33], DM2F [12], FFA [42], MSBDN [13], UHD [61], AECR [55], Dehamer [17], and DehazeFormer [49]), and three video dehazing methods (*i.e*., EVD [27], VDH [46], and CG-IDN [25]). We also compare several video restoration methods, including FastDVD [50], EDVR [53], NCFL [19], and BasicVSR++ [7]. On REVIDE, we compare MAP-Net with state-of-the-art methods of [59].

Implementation details. We use the AdamW optimizer and the polynomial scheduler. The initial learning rate is set as $2{\times}10^{-4}$. The total number of iterations is 40K. The batch size is eight, and the patch size of input video frames is 256×256. The weights λ_{phy} and λ_{flow} in Eq. (6) are empirically set as 0.2 and 0.04.

5.2. Comparisons with State-of-the-Art Methods

Quantitative comparison. Table 1 summarizes the quantitative results of our network and compared methods on HazeWorld. From these quantitative results, we can find that our method outperforms other baselines by a significant margin. Specifically, among all compared methods, BasicVSR++ [7] and DehazeFormer [49] achieve the best PSNR score of 26.06 and the best SSIM score of 0.9286, respectively. More importantly, our MAP-Net has a PSNR improvement of 1.06 dB over BasicVSR++, while our method has an SSIM gain of 0.0063 over DehazeFormer.

Table 2 shows the PSNR and SSIM of our network and state-of-the-art methods on REVIDE. Among all compared methods, NCFL [19] has the best PSNR (23.63 dB) and the best SSIM (0.8925). And our method further improves the PSNR from 23.63 dB to 24.16 dB and the SSIM from 0.8925 to 0.9043.

Qualitative comparison. Fig. 7 and Fig. 8 visually compare dehazed results produced by our network and state-of-the-art methods on video frames from HazeWorld and REVIDE. Apparently, compared methods often tend to introduce color distortion, darken several areas, or preserve some haze in their dehazed results, while our MAP-Net can effectively remove haze, avoid color distortion, and better recover the underlying clear frames. And the predicted haze-free results produced by our method are closest to ground truths shown in the last column of Fig. 7 and Fig. 8.

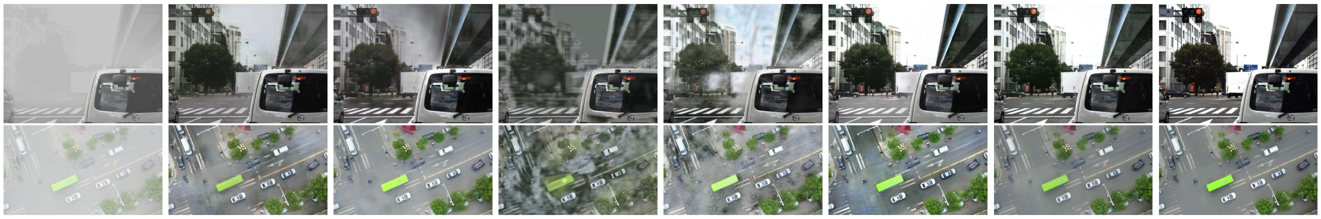


Figure 7. **Visual results on our HazeWorld.** Our method clearly removes haze and keeps more details. “GT” denotes the ground truth.

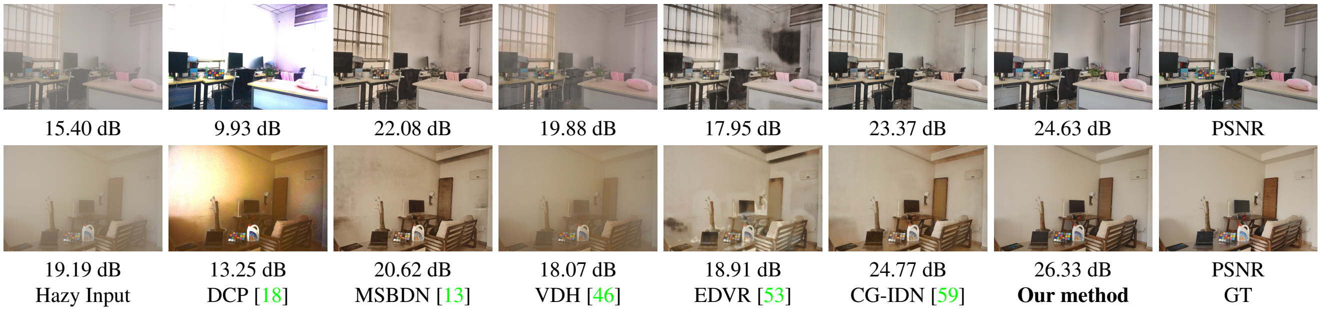


Figure 8. **Visual results on REVIDE [59].** Our method produces dehazed results with less haze artifact and color distortion.

Table 3. **Comparison of downstream effectiveness.** VPQ, mIoU, RMSE, and J&F are metrics for panoptic segmentation, semantic segmentation, depth estimation, and object segmentation.

Method	Hazy	MSBDN [13]	Dehamer [17]	EDVR [53]	BasicVSR++ [7]	Ours	GT
Cityscapes-VPQ ↑	21.7	40.2	43.4	47.2	45.9	**48.5**	56.5
Cityscapes-mIoU ↑	51.8	47.0	54.1	64.8	63.6	**66.2**	75.4
DDAD-RMSE ↓	21.21	14.99	15.01	15.26	14.89	**14.71**	14.36
DAVIS-J&F ↑	76.3	79.2	79.4	79.3	79.4	**80.0**	81.3

Moreover, we also compare our network against state-of-the-art methods on real-world hazy videos, and the results are shown in Fig. 9. From these visual results, we can find that existing methods tend to darken many areas, or maintain some haze. Compared to these methods, our network has a higher visual quality and less color distortion; see the last column of Fig. 9.

Applications. Our video dehazing method benefits several downstream applications, including video panoptic segmentation [21], object segmentation [41], depth estimation [16], and image semantic segmentation [10]. To verify this, we choose four different methods [16, 21, 40, 60] for corresponding downstream application validation, and obtain results on the input hazy videos, the dehazed videos, and the underlying haze-free videos. Table 3 reports the quantitative results. Apparently, the dehazed videos produced by different methods improve the downstream application performance compared to the original hazy videos. Notably, our method can better facilitate downstream applications than other representative dehazing methods.

5.3. Ablation Studies

We conduct a series of ablation studies on our HazeWorld dataset to analyze the effectiveness of major components of our network.

Table 4. **Ablation studies of our MPG and MSR modules.**

	(a)	(b)	(c)	Our method
Basic	✓	✓	✓	✓
MPG		✓		✓
MSR			✓	✓
PSNR	25.37	26.24	26.38	**27.12**
SSIM	0.9087	0.9171	0.9292	**0.9349**

Ablation studies of two major modules. We start by constructing a basic model (denoted as “Basic”), which has the only scene decoder, and STDA only considers one previous frame for temporal alignment. Then, we gradually add two proposed modules, *i.e.*, MPG and MSR. Table 4 compares their results. Compared to “Basic”, the MPG module and the MSR module have a PSNR improvement of 0.87 dB and 1.01 dB, respectively, due to the physical model-based priors at the MPG module, and multiple temporal haze clues at the MSR module. Moreover, combining both MPG and MSR modules together into our method can further improve our video dehazing performance.

Effectiveness of guidance information in our MSR module. As shown in Fig. 4, our MPG module learns prior guidance features, *i.e.*, $\mathcal{J}$ and $\mathcal{P}$, to guide the STDA block to align video frames and the GMRA block to aggregate features in our MSR module for video dehazing. Therefore, we conduct an ablation study to evaluate the effectiveness of the guidance information for the STDA block and the GMRA block, respectively. We achieve this by building three baseline networks: (1) we remove guidance information from both the STDA block and the GMRA block; (2) we only use the guidance at the STDA block; (3) we only utilize the guidance at the GMRA block. Note that we keep both

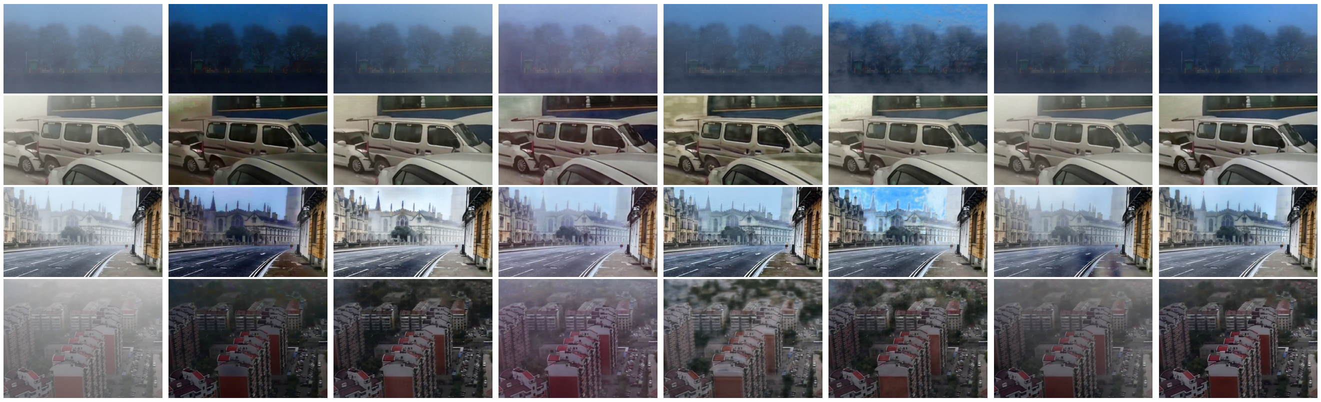


Figure 9. **Visual results on the real outdoor hazy videos.** Our method generates frames with more natural color and less haze remaining.

Table 5. **Effectiveness of the guidance at the STDA and GMRA blocks of the MSR module.**

STDA		✓		✓
GMRA			✓	✓
PSNR	26.38	26.92	26.61	**27.12**

Table 6. **Ablation studies of the number of ranges in MSR.**

(a) Discussion on the number of ranges.

#Range	1	2	3 (Ours)	4
PSNR	26.24	26.71	**27.12**	26.84

(b) Discussion on the temporal alignment.

Manner	1set	3sets-1range	Our method
PSNR	26.39	26.64	**27.12**

the STDA block and the GMRA block in our MSR module but only remove the guidance parts during the ablation. As shown in Table 5, our video dehazing performance is reduced if we remove the prior guidance information from either the STDA block or the GMRA block.

Discussion on different ranges of our MSR module. We perform two ablation study experiments on our MSR module: one is to discuss the number of space-time ranges used in our MSR module, while another experiment is to discuss how to leverage multiple adjacent frames when the number of ranges is fixed. Table 6a reports PSNR scores of our video dehazing with different numbers of ranges. We can observe that our PSNR is progressively improved when we increase the number of ranges from one to three. The reason is that a larger range value involves more temporal information for aligning video frames. However, when we further increase the range value from three to four, our video dehazing performance is reduced since the weights are shared for the STDA blocks, where a single STDA needs to tackle different ranges of temporal alignment. Hence, a larger range may introduce difficulty in the flow estimation, and thus we empirically set the number of frames/ranges as three.

Moreover, we further discuss how to use the adjacent video frames for temporal alignment when the number of frames is fixed as three. Here, we construct two baselines: (1) the 1st baseline (denoted as “1set”) is to change the three-set alignment in our method to only one set, which takes all three neighboring video frames. (2) the 2nd baseline (denoted as “3sets-1range”) is constructed by only using one adjacent frame in each set, *i.e.*, frame-by-frame alignment (three sets in total). As shown in Table 6b, our method has a better PSNR value than “1set” and “3sets-1range”, which indicates that considering three frames at multiple ranges incurs a better video dehazing performance.

6. Conclusion

This work designs a video dehazing framework via a multi-range temporal alignment network with physical prior. Two new techniques, a memory-based physical prior guidance module and a multi-range scene radiance recovery module, are formulated to effectively explore the physical haze priors and aggregate temporal information. We construct the first large-scale benchmark dataset for outdoor video dehazing, which enables us to evaluate the dehazing performance on various application scenarios and downstream tasks. In the end, the experimental results on both synthetic and real conditions demonstrate the superior of our framework against the recent state-of-the-art methods.

Limitations. Our method might not work well for videos with extremely heavy haze, and more prior knowledge is required. Also, though our method achieves superior performance and faster speed than many dehazing methods, it still cannot meet the real-time requirement.

Acknowledgment. This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: 14201321 and 14201620), the National Natural Science Foundation of China (Grant No. 61902275), Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), the National Key R&D Program of China (NO.2022ZD0160100), and Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

[1] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *CVPRW*, 2020. 2

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

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 3

[4] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 2016. 2

[5] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[6] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 2

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 6, 7, 8

[8] Chen Chen, Minh N Do, and Jue Wang. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *ECCV*, 2016. 2

[9] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *WACV*, 2019. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3, 7

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2

[12] Zijun Deng, Lei Zhu, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Qing Zhang, Jing Qin, and Pheng-Ann Heng. Deep multi-model fusion for single-image dehazing. In *ICCV*, 2019. 2, 6

[13] 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. 2, 6, 7, 8

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 4

[16] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 3, 7

[17] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, 2022. 2, 6, 7, 8

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 2010. 1, 2, 6, 7, 8

[19] Cong Huang, Jiahao Li, Bin Li, Dong Liu, and Yan Lu. Neural compression-based feature learning for video restoration. In *CVPR*, 2022. 2, 6

[20] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 4

[21] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 3, 7

[22] Jin-Hwan Kim, Won-Dong Jang, Jae-Young Sim, and Chang-Su Kim. Optimized contrast enhancement for real-time image and video dehazing. *Journal of Visual Communication and Image Representation*, 2013. 2

[23] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018. 2, 5

[24] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 3

[25] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 6

[26] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017. 2, 6

[27] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. End-to-end united video dehazing and detection. In *AAAI*, 2018. 1, 2, 6

[28] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 2018. 2

[29] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. Single image dehazing via conditional generative adversarial network. In *CVPR*, 2018. 2

[30] Zhuwen Li, Ping Tan, Robby T Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *CVPR*, 2015. 2

[31] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, 2022. 2

[32] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. In *ICML*, 2022. 2

[33] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, 2019. 2, 6

[34] Ye Liu, Liang Wan, Huazhu Fu, Jing Qin, and Lei Zhu. Phase-based memory network for video dehazing. In *ACM MM*, 2022. 2

[35] Ye Liu, Lei Zhu, Shunda Pei, Huazhu Fu, Jing Qin, Qing Zhang, Liang Wan, and Wei Feng. From synthetic to real: Image dehazing collaborating with unlabeled real data. In *ACM MM*, 2021. 2

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3

[38] Earl J McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York*, 1976. 1

[39] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019. 3

[40] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 7

[41] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 3, 7

[42] 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

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

[44] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 2, 5

[45] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016. 2

[46] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *TIP*, 2018. 1, 2, 6, 7, 8

[47] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 3

[48] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 3

[49] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *TIP*, 2023. 2, 6

[50] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *CVPR*, 2020. 6

[51] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 5

[53] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2, 6, 7, 8

[54] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020. 3

[55] 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. 6

[56] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 2

[57] Wenhan Yang, Jiaying Liu, and Jiashi Feng. Frame-consistent recurrent video deraining with dual-level flow. In *CVPR*, 2019. 2

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

[59] 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 *CVPR*, 2021. 1, 2, 6, 7

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 7

[61] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, and Xiuyi Jia. Ultra-high-definition image dehazing via multi-guided bilateral learning. In *CVPR*, 2021. 6

[62] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *TPAMI*, 2021. 3