Driving-Video Dehazing with Non-Aligned Regularization for Safety Assistance

Junkai Fan[1], Jiangwei Weng[1], Kun, Wang[1], Yijun Yang[2], Jianjun Qian[1], Jun Li[1,*], and Jian Yang[1*]

[1]PCA Lab[†], Nanjing University of Science and Technology, China

[2]The Hong Kong University of Science and Technology (Guangzhou)

{junkai.fan, wengjiangwei, kunwang, csjqian, junli, csjyang}@njust.edu.cn

yang018@connect.hkust-gz.edu.cn

Abstract

Real driving-video dehazing poses a significant challenge due to the inherent difficulty in acquiring precisely aligned hazy/clear video pairs for effective model training, especially in dynamic driving scenarios with unpredictable weather conditions. In this paper, we propose a pioneering approach that addresses this challenge through a non-aligned regularization strategy. Our core concept involves identifying clear frames that closely match hazy frames, serving as references to supervise a video dehazing network. Our approach comprises two key components: reference matching and video dehazing. Firstly, we introduce a non-aligned reference frame matching module, leveraging an adaptive sliding window to match high-quality reference frames from clear videos. Video dehazing incorporates flow-guided cosine attention sampler and deformable cosine attention fusion modules to enhance spatial multi-frame alignment and fuse their improved information. To validate our approach, we collect a GoProHazy dataset captured effortlessly with GoPro cameras in diverse rural and urban road environments. Extensive experiments demonstrate the superiority of the proposed method over current state-of-the-art methods in the challenging task of real driving-video dehazing. Project page.

1. Introduction

Haze significantly degrades visual quality, leading to challenges such as limited visibility and low contrast. This deterioration adversely affects high-level visual tasks crucial for safety in autonomous driving [28], including object detection [19], semantic segmentation [50], and depth estimation [54]. The degradation of haze effect can be expressed

---

(a) Temporal-spatial misalignment due to avoiding pedestrians and vehicles.

(i) Collecting scenes in hazy weather

(ii) Collecting scenes in clear weather

❌ *W/o Matched video frame pairs*   ✅ *We matched video frame pairs*

(b) Temporal misalignment in real-world scene hazy video pairs

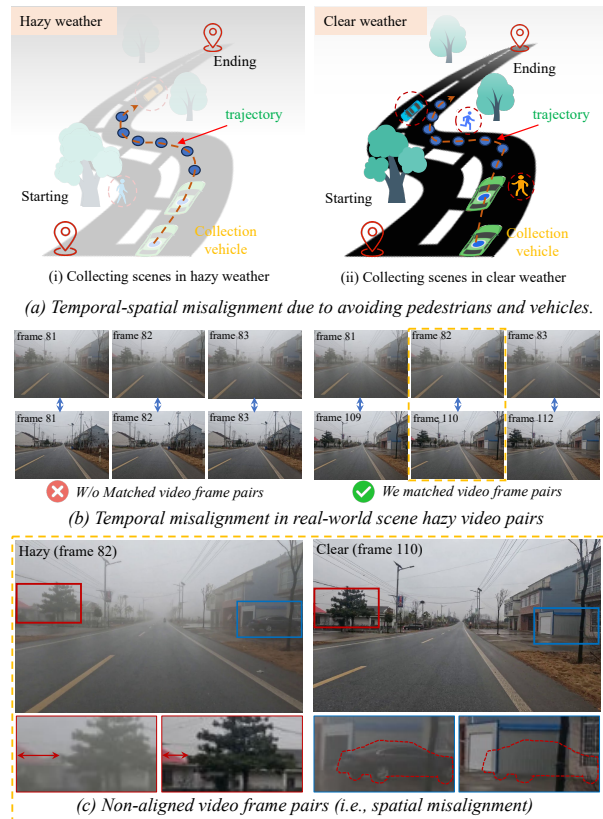*(c) Non-aligned video frame pairs (i.e., spatial misalignment)*

Figure 1. Spatial and temporal misalignments in real driving hazy/clear video pairs due to inconsistent driving speeds, different driving paths and moving objects.

through an atmospheric scattering model [39, 43] :

$$I(x) = J(x)t(x) + A_\infty(\lambda)(1 - t(x)), \quad (1)$$

where $I(x)$ and $J(x)$ represent the hazy image and the clear image at a pixel position x, respectively. A_∞ denotes the infinite airlight. The transmission map is defined as $t(x) = e^{-\beta(\lambda)d(x)}$, where $d(x)$ and $\beta(\lambda)$ signify the scene depth and the scattering coefficient associated with the wavelength of light λ, respectively. Although image/video dehazing [20, 59] has been extensively studied

over many years, there has been limited research on driving-video dehazing as dynamic driving scenarios with unpredictable weather conditions results in the inherent difficulty in acquiring precisely aligned ground truth (GT) videos for model training in Fig. 1 (a).

Here, we introduce a new paradigm for data collection that involves capturing driving videos under both hazy and clear conditions within the same scenes. This relaxes effectively the stringent requirement of strictly aligned GT. To assess its efficacy, we compile a GoPro-Hazy dataset, effortlessly recorded using GoPro cameras across various rural and urban road environments. Despite the ease of collecting hazy/clear video pairs, two challenges persist: temporal misalignment and spatial misalignment in the hazy and clear video pairs. Firstly, inconsistent driving speeds result in temporal misalignment. For example, as illustrated in Fig. 1 (b), frame 81 in the hazy video corresponds to frame 109, not frame 81 in the clear video. Secondly, distinct driving paths and moving objects contribute to spatial misalignment. As depicted in Fig. 1 (c), the car in the hazy video is not aligned with the corresponding scene.

To address spatial and temporal misalignment, this paper introduces an innovative driving-video dehazing method incorporating a non-aligned regularization learning approach. The fundamental concept involves identifying clear frames that closely match hazy frames as references to supervise a video dehazing network. Our method comprises two key components: reference matching and video dehazing. To enhance the quality of references, we introduce a Non-aligned Reference Frames Matching (NRFM) module, which pairs the input hazy frame with the clearest frame that most closely resembles the scene. Subsequently, we present a video dehazing model featuring a Flow-guided Cosine Attention Sampler (FCAS) module and a Deformable Cosine Attention Fusion (DCAF) module. FCAS utilizes pretrained coarse optical flow for multi-scale cosine attention sampling, improving offset accuracy and aligning multiple frames. Unlike the 'warp' operation relying on precise optical flow, cosine attention sampling achieves more accurate offset learning using coarse optical flow. DCAF aggregates multi-frame features by combining deformable convolution (DConv) [10] with a large receptive field and leveraging the robustness of cosine similarity for correlation computation. Prior to inputting the video dehazing model, we employ an image dehazing network to pre-remove haze from each frame. Our contributions can be summarized as follows:

- To our best knowledge, we are the first to propose a non-aligned regularization strategy for the real driving-video dehazing task. Its key idea is to selectively identify high-quality reference frames from the non-aligned clear video for supervision, reducing reliance on ground truth.
- We introduce a cutting-edge video dehazing network equipped with flow-guided cosine attention sampler and

deformable cosine attention fusion, effectively handling large motion in driving scenes.
- We provide a real-world video hazy dataset, which includes 27 non-aligned hazy/clear video pairs, totaling 4256 matched hazy/clear frame pairs. These pairs were collected manually using GoPro cameras in various real scenes (*i.e.*, countryside and urban roads).

2. Related Work

Image dehazing. Early approaches to single-image dehazing primarily concentrated on integrating atmospheric scattering models [39] with various priors [2, 16, 20, 75]. In contrast, later advancements in the field showcased superior performance through deep learning techniques, leveraging extensive datasets of hazy/clear images [1, 25]. These methods employ deep neural networks to either learn physical model parameters [13, 15, 26, 27, 30, 31, 34, 37, 42, 44, 58, 68, 69] or directly capture the mapping between hazy and clear images [9, 12–14, 29, 36, 46, 48, 52, 65]. For the latter category, recent works have introduced more sophisticated network structures, including transformer networks [18, 47, 54, 56]. However, these approaches heavily rely on aligned synthetic data for supervised learning, leading to suboptimal dehazing performance in real-world scenarios. To tackle this limitation, some studies have proposed domain-adaptive techniques [7, 45, 51, 59, 66] and unpaired dehazing models [6, 62, 63, 72] tailored for real scenes. Despite these efforts, when applying image dehazing models to videos, the outcomes often exhibit discontinuities due to the disregard for temporal information.

Video dehazing. Compared to single-image dehazing, video dehazing offers advantages by leveraging temporal cues from neighboring frames. Early approaches focused on enhancing temporal consistency in dehazing results, achieved through the optimization of transmission maps [50] and the elimination of artifacts [4]. Some methods also addressed multiple tasks concurrently, such as depth estimation [32], detection [24], within hazy videos. Recently, Zhang *et al*. [71] collected a real indoor smoke video dataset with ground truth, named REVIDE, and introduced a confidence-guided and improved deformable network (CG-IDN) for video dehazing. Building upon REVIDE, Liu *et al*.[38] proposed a novel phase-based memory network designed to enhance video dehazing by integrating both phase and color memory information. Similarly, Xu *et al*. [60] introduced a memory-based physical prior guidance module that encodes prior-related features into long-term memory for video dehazing. Furthermore, certain video restoration methods [21, 64], demonstrate superior performance on the REVIDE dataset for adverse weather conditions. However, it's crucial to note that these approaches are primarily trained and evaluated in indoor smoke scenes. As a result, their effectiveness in addressing complex real-
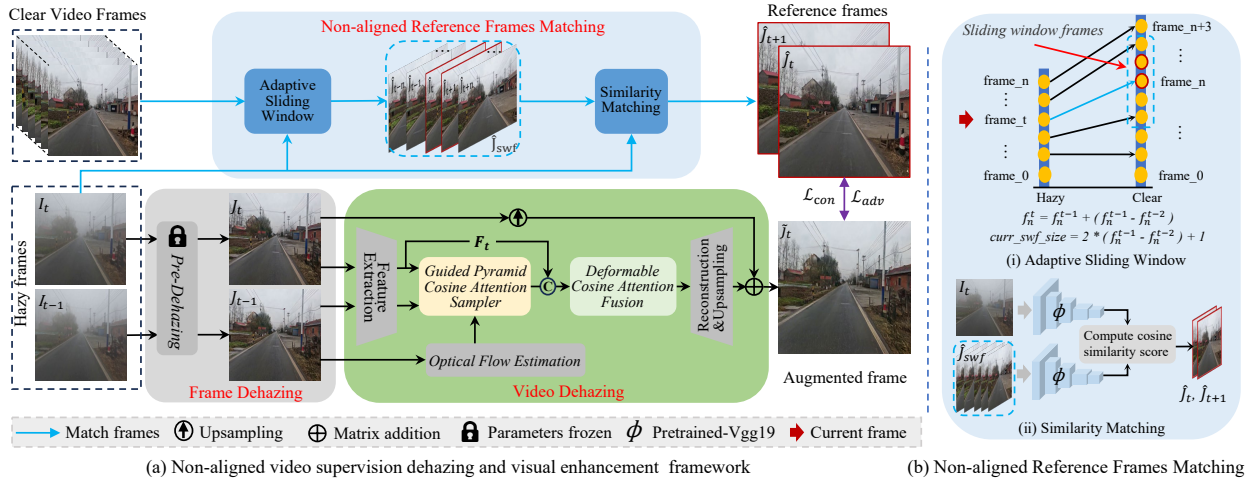
Figure 2. (a) The overall framework of our driving-video dehazing (DVD) comprising two crucial components: frame matching and video dehazing. This involves applying frame dehazing to proactively eliminate haze from individual frames. One significant benefit is is the effectiveness and efficiency of our method in training the video dehazing network using authentic driving data without requiring strict alignment, ultimately producing high-quality results. (b) The illustration depicts the matching process of non-aligned, clear reference frames through the utilization of an adaptive sliding window using feature cosine similarity. *Our input consists of two frames.*

world outdoor haze conditions remains limited.

Video alignment. The primary objective of alignment is to capture spatial transformations and pixel-wise correspondence between adjacent frames. Video-related tasks, like restoration and super-resolution, often face alignment challenges [5, 35]. Recent works rely on precise optical flow estimation [49] to align adjacent images/features [21, 22, 61]. Alternatively, some approaches leverage deformable convolution (DConv) [10] to learn feature alignment offsets [55, 57, 71]. Other methods [3, 33, 60, 70] employ attention mechanisms to combine optical flow and DConv for feature alignment. However, these alignment methods face two challenges: 1) obtaining accurate optical flow with pretrained models is difficult, and 2) DConv training is unstable under large motion conditions.

In comparison to the aforementioned supervised video dehazing methods [60, 71], our approach surpasses previous video dehazing models. This is achieved by training on non-aligned real-world hazy datasets and extracting effective features from clear and misaligned reference frames within the same scene. Furthermore, we introduce a Flow-guided Cosine Attention Sampler (FCAS) module, which more accurately aligns multi-frame features under inaccurate optical flow conditions by incorporating learnable multi-scale cosine attention sampling.

3. Methodology

Here, we present an innovative driving-video dehazing method illustrated in Fig. 2 (a). Initially, we introduce a Non-aligned Reference Frame Matching (NRFM) module, employing an adaptive sliding window that utilizes feature similarity to match high-quality reference frames for supervising the video dehazing network in subsection 3.1. Subse-

quently, we propose a video dehazing module that integrates a flow-guided cosine attention sampler and deformable cosine attention fusion. This integration aims to improve spatial multi-frame alignment and fuse the enhanced information from multiple frames in subsection 3.2. Before displaying them, we first pre-process the hazy frames.

For a given continuous hazy/clear video pair ($I = I_{[0:N]}, \widehat{J} = \widehat{J}_{[0:M]}$) with $N \leq M + 2$, we utilize an image dehazing method to pre-remove haze from each frame,

$$J_t = \mathcal{P}(I_t), \tag{2}$$

where $\mathcal{P}$ denotes an image dehazing network, and we employ the non-aligned supervision network [15]. So, the video pair is rewritten as ($J = J_{[0:N]}, \widehat{J} = \widehat{J}_{[0:M]}$).

Prioritizing frame dehazing offers two key advantages. First, easily acquiring non-aligned image pairs simplifies the training process with a large dataset, leading to high-quality pre-processing outcomes. Second, superior frame dehazing enhances the video dehazing stage's capability to learn pixel correlations among adjacent frames.

3.1. Non-aligned Reference Frame Matching

In this subsection, for the hazy video I, our main objective is to establish its corresponding clear and non-aligned reference frames derived from the clear video $\widehat{J}$ in Fig. 2 (b). These reference frames serve as supervision for the video dehazing network. Further, we curate a set of *non-aligned video pairs* characterized by temporal and spatial misalignments in Fig. 1 (b) and (c).

To solve temporal misalignment, we introduce a non-aligned reference frame matching module to match the clear reference frames in Fig. 2 (b). For each hazy frame I_t, we formally denote its corresponding sliding window clear
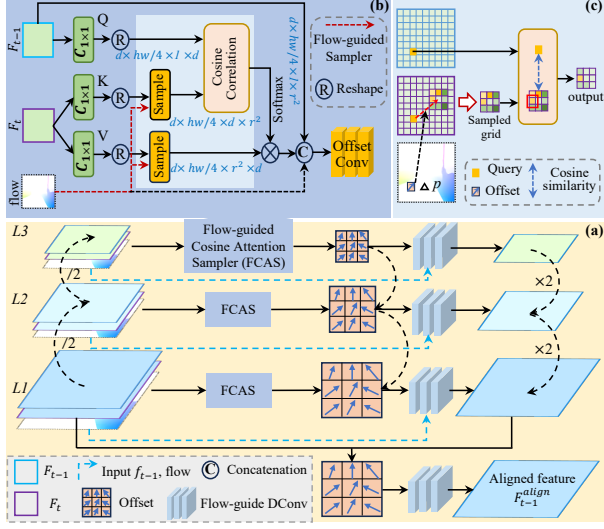
Figure 3. (a) Overview of guided pyramid cosine attention sampler (GPCAS). (b) The proposed FCSA module uses coarse optical flow sampling to enhance the receptive field for cosine correlation calculations. (c) Sampling and calculating cosine correlation.

frames as $\widehat{J}_{[i_s^t : i_e^t]}$, where i_s^t and i_e^t denote the starting and ending indexes, respectively. When $t = 0$, we initialize i_s^0 and i_e^0 as 0 and $\lceil (M - N)/2 \rceil$, respectively. To iteratively match clear reference frames, we define the iterated indexes at the t-th frame as:

$$i_s^t = i_s^{t-1} + (k^{t-1} - k^{t-2}), \qquad (3)$$

$$i_e^t = 2(k^{t-1} - k^{t-2}) + 1, \qquad (4)$$

where k^t represents the index of the most similar clear frame from $\widehat{J}_{[i_s^t : i_e^t]}$, determined by comparing their cosine similarity. The index is defined as:

$$k^t = \arg \min_{i_s^t \leq i \leq i_e^t} \left\{ d\left(\Phi(I_t), \Phi(\widehat{J}_i) \right) \right\}, \qquad (5)$$

where Φ denotes the VGG-16 [53] network. Consequently, we obtain the matching reference frames $\widehat{J}_{k^t}$ and $\widehat{J}_{k^t+1}$ for the hazy frame I_t. The overall procedure of our NRFM is outlined in **Algorithm 1**.

Multi-frames Reference Loss. In addition to temporal misalignment, our collected data also exhibits pixel and semantic misalignment in Fig. 1 (c). To tackle spatial misalignment, we devise a multi-frame reference loss to ensure feature consistency between the video dehazing result $\widetilde{J}_t$, and the reference frames $\widehat{J}_{k^t}$ and $\widehat{J}_{k^t+1}$. Based on the contextual loss [40] and cosine distance, our multi-frame reference loss is formulated as

$$\mathcal{L}_{\text{mfr}}(\widetilde{J}_t, \widehat{J}_{k^t}, \widehat{J}_{k^t+1}) = \sum_{l=1}^{5} d\left(\Phi^l(\widetilde{J}_t), \Phi^l(\widehat{J}_{k^t}) \right) + \\ \sum_{l=1}^{5} d\left(\Phi^l(\widetilde{J}_t), \Phi^l(\widehat{J}_{k^t+1}) \right), \quad (6)$$

where $d(\cdot, \cdot)$ is the cosine distance between $\widetilde{J}_t$ and $\widehat{J}_{k^t}$. $\Phi^l(\widetilde{J}_t)$ and $\Phi^l(\widehat{J}_{k^t})$ represent the feature maps extracted
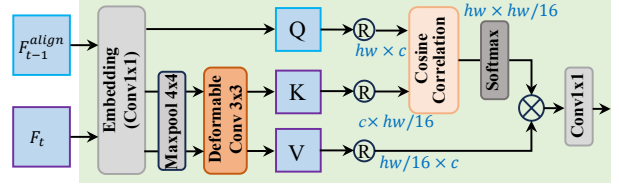

Figure 4. Overview of proposed DCAF. Enhancing cosine correlation for pixel misalignment robustness by expanding the receptive field with DConv, thereby improving cosine fusion performance.

Algorithm 1: NRFM (default $N \leq M + 2$)

Input: hazy video: $I_{[0:N]}$, clear video: $J_{[0:M]}$

Output: $[\widehat{J}_k, \widehat{J}_{k+1}]$

1 Initialize: $i_s^0 = 0$, $i_e^0 = \lceil (M - N)/2 \rceil$, $\widehat{J}_k = []$ and $\widehat{J}_{k+1} = []$;

2 **for** $t = 0, ..., N$ **do**

3   Compute the index k^t by Eq. (5);

4      $\widehat{J}_k = [..., \widehat{J}_k, \widehat{J}_{k^t}]$;

5      $\widehat{J}_{k+1} = [..., \widehat{J}_{k+1}, \widehat{J}_{k^t+1}]$;

6   Update i_s^t and i_e^t by Eqs. (3) and (4);

7 **end**

---

from the l-th layer of the VGG-16 network with inputs $\widetilde{J}_t$ and $\widehat{J}_{k^t}$, respectively. k^t is the matching index of the clear reference frame of the t-th hazy frame.

3.2. Video Dehazing

In video tasks, previous studies [55, 67, 70] have revealed the significance of a larger receptive field. This attribute proves beneficial for aligning and fusing adjacent frames, as it extends the search range and facilitates the learning of pixel correlations between neighboring frames. The prevailing approaches often involves using optical flow for warping alignment [3, 70, 74]. However, these methods are limited by optical flow precision, especially when dealing with blurry images after pre-dehazing.

Motivated by these observations, we propose a novel Flow-guided Cosine Attention Sampler (FCAS) module. This module leverages coarse optical flow for sampling, thereby expanding the receptive field for cosine correlation calculations. This augmentation enhances computational accuracy and yields superior alignment results, as depicted in Fig. 3. Additionally, we extend this concept to introduce a Deformable Cosine Attention Fusion (DCAF) module, illustrated in Fig. 4. The DCAF module employs deformable convolutions (DConv) to broaden sampling receptive fields, capturing long-term dependencies and thereby improving feature aggregation across multiple frames.

3.2.1 Flow-guided Cosine Attention Sampler

In Fig. 3 (b), our FCAS module aims to align the features of the previous frame F_{t-1} with those of the current frame

F_t. The FCAS module produces the offset between adjacent frame features $[F_{t-1}, F_t] \in \mathbb{R}^{C \times H \times W}$, where C, H, and W denote the channel, height, and width of the features, respectively. Additionally, an optical flow $O_{t-1 \to t}$ is learned to capture pixel-to-pixel correspondence from the previous frame to the current frame.

Specifically, F_{t-1} and F_t are derived from a feature extraction network applied to the pre-dehazing results $[J_{t-1}, J_t]$. The optical flow $O_{t-1 \to t} \in \mathbb{R}^{2 \times H \times W}$ is obtained by fine-tuning SpyNet [49], denoted as ϕ_{spy}, during training. The flow offset map at each position $p = (x, y)$ in I_{t-1} is mapped to its estimated correspondence in I_t as $p' = (x + u, y + v)$, which is defined as

$$\Delta p = (u, v) = \phi_{spy}(I_{t-1}, I_t)(x, y). \tag{7}$$

The set of sampled grid coordinates is expressed as

$$\Omega(p')_k = \{p' + e \mid e \in \mathbb{Z}^2, ||e||_1 \leq (k-1)/2\}, \tag{8}$$

where k represents the sampling kernel size and $\mathbb{Z}^2$ denotes a two-dimensional space. Linear projected query vectors $Q_{x,y} = F_{t-1}W^q$, key vectors $K_{x,y} = F_t W^k$, and value vectors $V_{x,y} = F_t W^v$ at coordinate $p = (x, y)$ of F_{t-1} and F_t are defined using the parameters W^q, W^k, and $W^v \in \mathbb{R}^{C \times d}$, where d is the dimension of the projected vector.

Fig. 3 (c) illustrates the use of the coarse $O_{t-1 \to t}$ to guide learnable sampling from $K_{x,y}$ and $V_{x,y}$, expanding the receptive field for cosine correlation calculations to enhance accuracy. Within the sampled grid coordinates, the sampling key and value elements are described as

$$\{K_{i,j}, V_{i,j} \mid (i, j) \in \Omega(p')_k\} = \mathcal{S}(K_{x,y}, V_{x,y}), \tag{9}$$

where $\mathcal{S}$ denotes the interpolation sampling. The cosine attention $F_{\text{attn}} \in \mathbb{R}^{HW/4 \times 1 \times k^2}$ is then computed by

$$F_{\text{attn}} = \sum_{(i,j) \in \Omega(p')_k} F_{\text{softmax}} \left(\frac{Q_{x,y}^T K_{i,j}}{|Q_{x,y}||K_{i,j}|\sqrt{d}} \right) V_{i,j}, \tag{10}$$

where d is the dimension of the projected vector. Finally, the output offset is computed as

$$o_{t-1 \to t} = \text{Conv}\left(\text{Cat}(F_{t-1}, F_{attn}, O_{t-1 \to t})\right), \tag{11}$$

where Cat represents the concatenation operation, and $o_{t-1 \to t}$ is the offset map between F_{t-1} and F_t.

3.2.2 Deformable Cosine Attention Fusion

Similar to the central concept discussed in Section 3.2.1, enhancing the accuracy of cosine correlation calculation involves expanding the receptive field. However, a distinction arises in DCAF (refer to Fig. 4), where we broaden the receptive field using deformable convolution (DConv).

To fully leverage the spatial cues from multiple frames, the DCAF module is employed to fuse the aligned feature F_{t-1}^{align} with the current frame feature F_t to achieve further alignment. Initially, we transform F_{t-1}^{align} and F_t to compute the embedding query Q_{t-1}^{align}, key K_t, and value V_t through convolutional operations with a 1×1 kernel size, denoted by C_1. Subsequently, the key K_t and value V_t undergo downsampling via a 4×4 maxpooling operation, denoted by $\mathcal{M}$. They are computed by

$$\widetilde{Q}_{t-1}^{\text{align}} = \mathcal{M}(C_1(Q_{t-1}^{\text{align}})), \tag{12}$$

$$\widetilde{K}_t = \text{DConv}(\mathcal{M}(C_1(K_t))), \tag{13}$$

$$\widetilde{V}_t = \text{DConv}(\mathcal{M}(C_1(V_t))). \tag{14}$$

Next, we use the Eq. (10) to calculate the cosine correlation, and obtain the fused feature $F_{\text{fusion}} \in \mathbb{R}^{C \times H \times W}$.

3.3. Training Loss

For frame dehazing, we exclusively utilize the pre-trained NSDNet [15], please refer to its training loss for details. Now, let's focus on elucidating the training loss for video dehazing, which is expressed as follows:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{mfr}} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{cr}}, \tag{15}$$

$\mathcal{L}_{\text{adv}}$ represents the adversarial loss [17], and $\mathcal{L}_{\text{mfr}}$ corresponds to the multi-frames reference loss as defined in Eq. (6). Since we lack the ground truth for the aligned feature F_{t-1}^{align}, we optimize the guided pyramid cosine attention sampler (GPCAS) module by using the current frame feature F_t as the label. Our objective is to minimize the discrepancy between F_{t-1}^{align} and F_t, expressed as $\mathcal{L}_{\text{align}} = ||F_{t-1}^{\text{align}} - F_t||_1$. Inspired by [11], we introduce a self-supervised temporal consistency regularization to ensure the consistency (i.e., color and brightness) of pixels between consecutive frames. It can be formulated as:

$$\mathcal{L}_{\text{cr}} = ||M \odot (\mathcal{W}_{t \to t-1}(\widetilde{J}_t, \mathcal{O}_{t \to t-1}) - \widetilde{J}_{t-1}||_1, \tag{16}$$

where M is the occlusion map, $\mathcal{W}$ represents the flow-based image warp [49] for pixel alignment based on optical flow $\mathcal{O}_{t \to t-1}$, and $\widetilde{J}_{t-1}$ is the previous output frame.

4. GoProHazy and DrivingHazy Datasets

4.1. Collection Details

Camera Parameters Setting. We utilized a GoPro 11 camera with anti-flicker set to 60Hz, video output resolution at 1920x1080, frames per second (FPS) set to 30, and default focal length range of 19-39mm.

Collection Settings. Firstly, as shown in Fig. 5 (a), we use an electric vehicle to collect the GoProHazy dataset, ensuring controlled speed for higher-quality non-aligned

Data Settings	Methods	Data Type	GoProHazy		DrivingHazy (NoRef)			InternetHazy (Only testing)			Params (M)	Flops (G)	Ref.
			FADE ↓	NIQE ↓	FADE ↓	NIQE ↓	Votes ↑	FADE ↓	NIQE ↓	Votes ↑			
Unpaired	DCP [20]	Image	0.9835	5.8309	0.9692	5.6799	-	0.9223	6.4744	-	-	-	CVPR'09
	RefineNet [72]	Image	1.5694	5.3693	1.1837	5.5500	-	1.1801	5.8742	-	11.38	75.41	TIP'21
	CDD-GAN [6]	Image	1.1942	4.9787	1.4423	5.0349	-	1.2120	5.1049	-	29.27	56.89	ECCV'22
	D^4 [63]	Image	1.9272	5.7865	1.8658	5.6864	-	1.3277	6.2150	-	**10.70**	**2.25**	CVPR'22
Paired	PSD [7]	Image	1.0529	6.0010	0.9672	5.3520	-	0.9275	5.2187	-	33.11	182.5	CVPR'21
	RIDCP [59]	Image	0.8010	4.6640	1.1077	4.3889	0.315	0.9391	4.6610	0.265	28.72	182.69	CVPR'23
	PM-Net [38]	Video	1.1011	4.1211	0.9434	3.8944	0.220	1.1517	4.0590	0.150	151.20	5.22	ACMM'22
	MAP-Net [60]	Video	1.0611	4.2359	1.0440	4.2542	0.025	1.2130	5.3241	0.030	28.80	8.21	CVPR'23
Non-aligned	NSDNet [15]	Image	0.7996	4.1547	0.9348	4.0529	-	0.8934	4.3835	-	11.38	56.86	arXiv'23
	DVD (Ours)	Video	**0.7598**	**3.7753**	**0.8207**	**3.5825**	**0.440**	**0.8745**	**3.7480**	**0.555**	15.37	73.12	-

Table 1. Quantitative results on three real-world hazy video datasets. ↓ denotes the lower the better. ↑ denotes the higher the better. Due to PM-Net and MAP-Net rely on GT for training, we use $\mathcal{L}_{cx}$ to train them on GoProHazy dataset. Note that we only selected the latest dehazing methods (*i.e.*, RIDCP, PM-Net and MAP-Net) and our DVD for the user study. Moreover, DrivingHazy and InternetHazy were tested on dehazing models trained using GoProHazy and pre-trained dehazing models provided by the authors, respectively.

(a) GoProHazy's capture setting   (b) DrivingHazy's capture setting

Figure 5. Vehicles with different speeds for data collection.

hazy/clear video pairs at lower speeds (30 - 35 km/h). Secondly, as illustrated in Fig. 5 (b), we employ a car to capture the DrivingHazy dataset, testing performance under higher driving speeds (60 - 80 km/h) in a real-world environment. **Collection Method.** To collect non-aligned of hazy/clear video pairs, follow these steps:

- 1). As illustrated in Fig. 1 (a-i), we capture hazy videos in various scenes under hazy weather conditions.
- 2). In Fig. 1 (a-ii), to maintain consistent scene brightness, we choose overcast days with good visibility for capturing clear video pairs. Besides, to ensure the reference clear video matches the hazy scene, we align clear video capture with the starting point of the hazy videos.
- 3). Video cropping ensures consistent starting and ending points for collected hazy/clear video pairs.

4.2. Statistical Analysis

In Fig. 6, within the GoProHazy dataset, urban roads dominate our scenes, with 40% exhibiting heavy haze and 47% moderate haze. Overall, 87% of scenes depict visibility below 100 meters. In the DrivingHazy dataset, real high-speed scenario videos increased to 21%, with hazy density mainly in the 0-50 meters visibility range, constituting 54% of the dataset. In summary, both the GoProHazy and DrivingHazy datasets predominantly feature urban road scenarios, with hazy density concentrated within a 0-100m visibility range.

5. Experimental Results

We validate the effectiveness of our proposed method by experimenting with three real-world hazy video datasets.

(a) GoProHazy dataset (4256 frames)

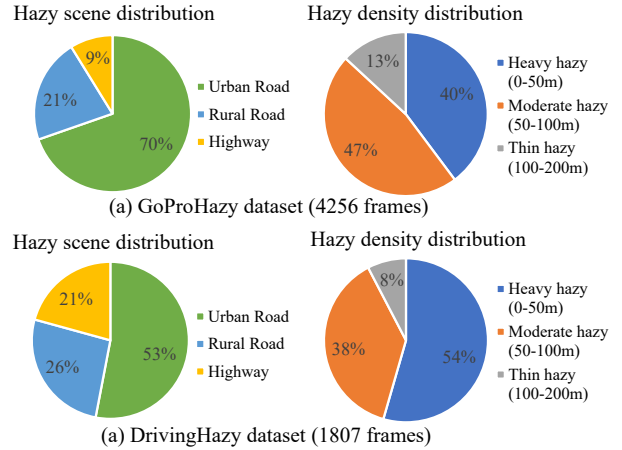(a) DrivingHazy dataset (1807 frames)

Figure 6. Statistical analysis of hazy scenes and density distribution in the GoProHazy and DrivingHazy datasets.

To assess its performance, we compare our method against state-of-the-art image and video dehazing techniques. Additionally, we conduct three ablation studies to substantiate the efficacy of our proposed core module. *Note that REVIDE dataset experiment, more visual results, ablation studies and video demo are provided in supplementary material.*

5.1. Experimental Setting

Three real-world hazy video datasets. One of these datasets is *GoProHazy*, where videos were captured using a GoPro camera under hazy and clear conditions. The recordings were made at the starting and ending points of the same road, with a total of 22 training videos (3791 frames) and 5 testing videos (465 frames). Each hazy video in the dataset is paired with a corresponding clear reference video, and the footage was obtained by driving an electric vehicle. In contrast, *DrivingHazy* was collected by driving a car at a relatively high speed in real hazy conditions. It comprises 20 testing videos (1807 frames), providing a unique perspective on hazy conditions during fast-paced driving. Moreover, we curated two distinct sets of hazy videos, contributing to the creation of the *InternetHazy*. This dataset, com-
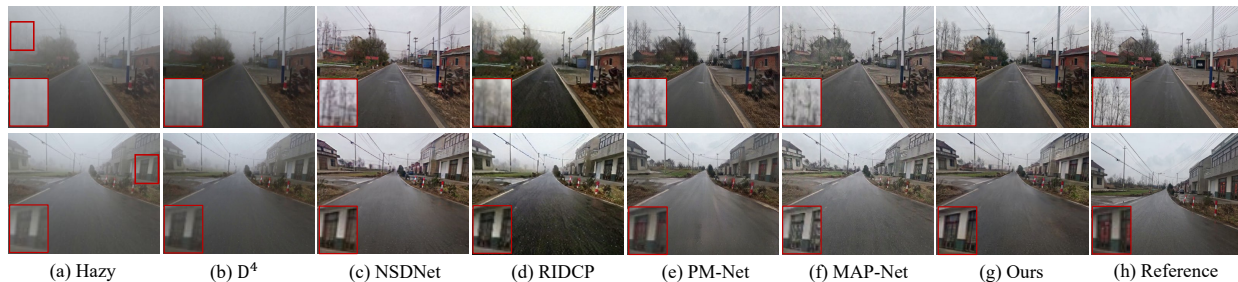
(a) Hazy   (b) D^4 (c) NSDNet   (d) RIDCP   (e) PM-Net   (f) MAP-Net   (g) Ours   (h) Reference

Figure 7. Comparison of video dehazing results on GoProHazy. Our method effectively removes distant haze.

(a) Hazy   (b) CDD-GAN   (c) D^4 (d) RIDCP   (e) NSDNet   (f) PM-Net   (g) MAP-Net   (h) Ours
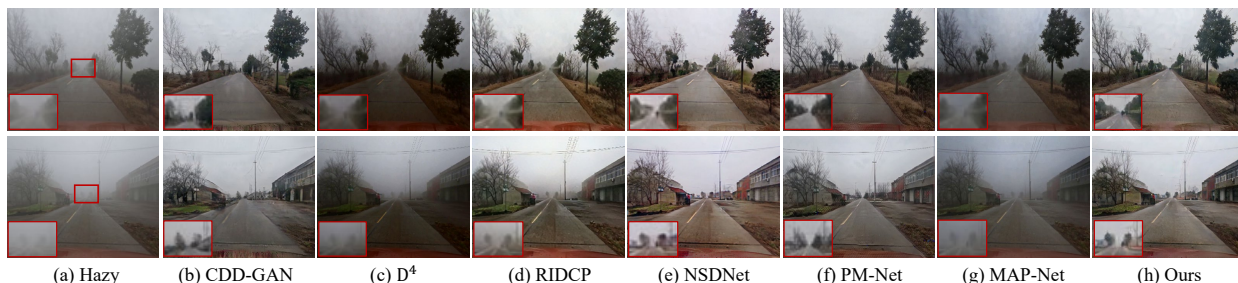
Figure 8. Testing results on DrivingHazy. Our method can perform dehazing in real driving environments while preserving the brightness.

prising 328 frames, features hazy videos with distributions distinct from those found in *GoProHazy*. It enriches our study by introducing diverse hazy scenarios for analysis.

Implementation details. In training processing, we use ADAM [23] optimizer with default parameter ($\beta_1 = 0.9$, $\beta_2 = 0.99$) and MultiStepLR scheduler. The initial learning rate is set as 1×10^{-5}. The batch size is 1, and the image size of input frames is 256×256. Our model was trained for 95K iterations by Pytorch with two NVIDIA RTX 3090 GPUs.

5.2. Main Results

Quantitative comparison. In Table. 1, our method outperforms SOTAs performance in terms of FADE [8] and NIQE [41] across all collected datasets. Specifically, on the GoProHazy dataset, our approach achieves the highest FADE score of 0.7598 and the best NIQE score of 3.7753, surpassing previous SOTA methods. Notably, our method exhibits a FADE improvement of 0.0412 over RIDCP and an NIQE gain of 0.3458 over PM-Net.

On the DrivingHazy dataset, our method achieves a FADE improvement of 0.1227 and a NIQE gain of 0.3119 over PM-Net, the leading competitor. Evaluating the generalization performance of our proposed DVD on the Internet-Hazy dataset without retraining or fine-tuning, our method consistently outperforms other approaches, solidifying its position as the top-performing model for generating dehazing results across diverse datasets.

In summary, our method surpasses supervised counterparts, leveraging non-aligned regularization. Unlike supervised approaches requiring pixel-wise alignment, our method excels by imposing robust constraints, such as obtaining image pairs within similar scenes to ensure a consistent distribution of clear and hazy images. Compared to

Mehods	wo / NRFM (unpaired)	Ours (misaligned)
FADE ↓	0.9204	**0.7598**
NIQE ↓	3.9729	**3.7753**

Table 2. Ablation study for our NRFM on GoProHazy.

unpaired competitors like D^4, our approach applies stronger constraints, leading to more effective dehazing results.

Visual comparison. The dehazing visualizations in Fig. 7 highlight the performance of our approach. Overall, our method exhibits superior brightness and texture details compared to other state-of-the-art (SOTA) techniques. Notably, D^4 and RIDCP fail to eliminate distant haze, with RIDCP additionally displaying color distortion. While PM-Net and MAP-Net successfully clear distant haze, they compromise on texture details, resulting in blurred images. Figs. 8 and 9 showcase visualizations on the DrivingHazy and InternetHazy datasets. Despite their advancements, state-of-the-art dehazing methods share a common limitation—they struggle to effectively remove distant haze while preserving texture details and brightness in the images. Moreover, we validated the effectiveness of our method in the *user study* results presented in Table. 1.

5.3. Ablation Studies

Effect of NRFM. To assess the effectiveness of our proposed NRFM, we conducted experiments by excluding the NRFM module and training our video dehazing model in an unpaired setting, where clear reference frames were randomly matched. The results in Table 2 and Fig. 10 show a notable improvement in video dehazing by integrating our NRFM module. This enhancement is due to a more robust supervisory signal from non-aligned clear reference frames, distinguishing it from the unpaired setting.
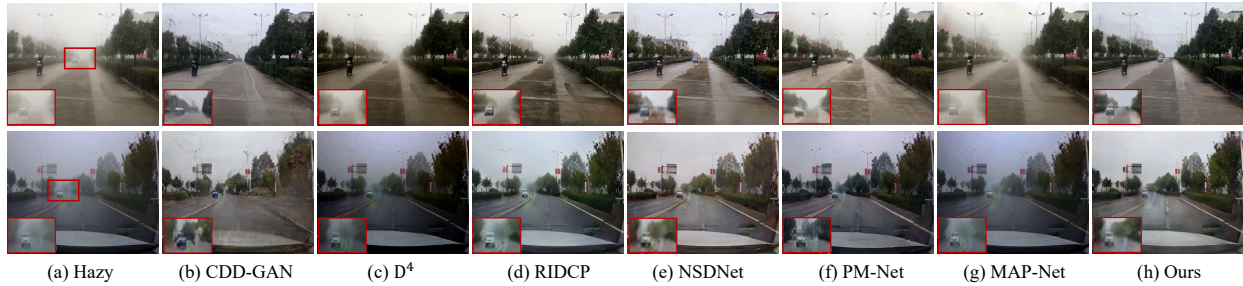
(a) Hazy | (b) CDD-GAN | (c) D⁴ | (d) RIDCP | (e) NSDNet | (f) PM-Net | (g) MAP-Net | (h) Ours

Figure 9. Testing with the pre-trained model provided by the authors on the InternetHazy dataset. Our method excels in dehazing.
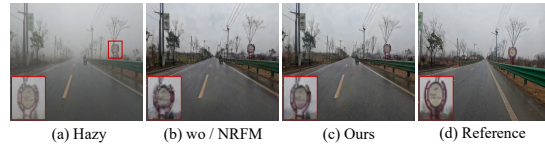

(a) Hazy | (b) wo / NRFM | (c) Ours | (d) Reference

Figure 10. Ablation visualization of our NRFM.

Methods	(i)	(ii)	(iii)	Our method
Basic	✓	✓	✓	✓
FCAS			✓	✓
DCAF		✓		✓
FADE ↓	0.8957	0.8869	0.8464	**0.8207**
NIQE ↓	3.9217	3.8495	3.6981	**3.5825**

Table 3. Ablation studies of FCAS and DCAF on DrivingHazy.

Effect of FCAS and DCAF. We conducted a series of experiments to validate the efficacy of the FCAS and DCAF modules on the DrivingHazy dataset. Initially, we developed a baseline video dehazing framework that comprised a frame dehazing module, a pyramid deformable convolution alignment module, and a non-local fusion module, referred to as model (i). This model was trained using adversarial loss ($\mathcal{L}_{adv}$) and multi-frames reference loss ($\mathcal{L}_{mfr}$). Subsequently, to assess the impact of the FCAS module, we integrated it into the pyramid deformable convolution alignment module, resulting in a comparative model (ii). Similarly, to evaluate the effectiveness of the DCAF module, we replaced the non-local fusion module with the deformable cosine fusion module, denoted as model (iii). Following this, we introduced our proposed modules (FCAS and DCAF) by replacing the pyramid deformable convolution alignment module and the non-local fusion module in the baseline model, forming our proposed method. The quantitative results are presented in Table 3, where our method exhibits the lowest FADE and NIQE values, indicating its excellent real-world video dehazing performance.

Additionally, the ablation results for different modules are visualized in Fig. 11. (a) displays the frame dehazing results used as input for video dehazing. (b), (c), (d), and (e) illustrate the visualized dehazing results for models (i), (ii), (iii), and our method, respectively. The dehazing results of models (i) and (iii) appear blurrier in comparison to our result in (e). Moreover, (c) exhibits reduced blurriness but lacks structural information of objects in the image.
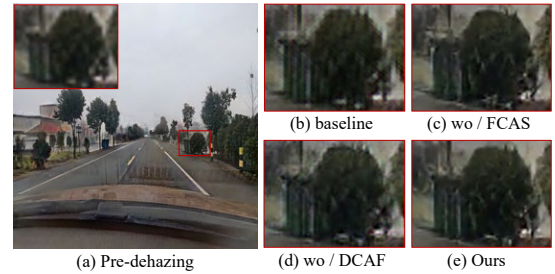

(a) Pre-dehazing | (b) baseline | (c) wo / FCAS | (d) wo / DCAF | (e) Ours

Figure 11. Ablation visualization of our FCAS and DCAF.

Kernel Size	(3×3)	(5×5)	(7×7)	(9×9)
FADE ↓	0.9626	0.9598	**0.7598**	0.9637
NIQE ↓	3.8307	3.8753	**3.7753**	3.8098

Table 4. Comparison with different kernel sizes on GoProHazy.

Effect of sampling kernel size. We conducted experiments using various kernel sizes to evaluate their influence on video dehazing outcomes. Due to computational constraints, we opted for kernel sizes of 3, 5, 7, and 9. Table 4 indicates that a 7×7 kernel size yields the most favorable results. Optimal sampling kernel size should account for motion magnitude between frames. A kernel size of 1×1 corresponds to a wrapping operation.

6. Conclusion

We introduce an innovative and effective video dehazing framework explicitly tailored for real-world driving scenarios with hazy videos. By leveraging non-aligned hazy/clear video pairs, we address the challenges of temporal and spatial misalignment through the incorporation of a non-aligned reference frame matching module. This module utilizes high-quality clear and misaligned reference frames, providing robust supervision for video dehazing. we enhance spatial multi-frame alignment and aggregation through the integration of flow-guided cosine attention sampler and deformable cosine attention fusion modules. Our framework's experimental results unequivocally demonstrate superiority over recent state-of-the-art methods, not only enhancing video dehazing but also promising improved visibility and safety in real driving scenarios.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pages 444–445, 2020. 2

[2] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1674–1682, 2016. 2

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5972–5981, 2022. 3, 4

[4] Chen Chen, Minh N Do, and Jue Wang. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 576–591. Springer, 2016. 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 (CVPR)*, pages 6286–6295, 2018. 3

[6] Xiang Chen, Zhentao Fan, Pengpeng Li, Longgang Dai, Caihua Kong, Zhuoran Zheng, Yufeng Huang, and Yufeng Li. Unpaired deep image dehazing using contrastive disentanglement learning. In *European Conference on Computer Vision (ECCV)*, pages 632–648. Springer, 2022. 2, 6

[7] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7180–7189, 2021. 2, 6

[8] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing (TIP)*, 24(11):3888–3901, 2015. 7

[9] Xiaofeng Cong, Jie Gui, Kai-Chao Miao, Jun Zhang, Bing Wang, and Peng Chen. Discrete haze level dehazing network. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, pages 1828–1836, 2020. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 764–773, 2017. 2, 3

[11] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17622–17631, 2022. 5

[12] Qili Deng, Ziling Huang, Chung-Chi Tsai, and Chia-Wen Lin. Hardgan: A haze-aware representation distillation gan for single image dehazing. In *European conference on computer vision (ECCV)*, pages 722–738. Springer, 2020. 2

[13] 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 *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 2453–2462, 2019. 2

[14] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10729–10736, 2020. 2

[15] Junkai Fan, Fei Guo, Jianjun Qian, Xiang Li, Jun Li, and Jian Yang. Non-aligned supervision for real image dehazing. *arXiv preprint arXiv:2303.04940*, 2023. 2, 3, 5, 6

[16] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems (NeurIPS)*, 27, 2014. 5

[18] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5812–5820, 2022. 2

[19] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15283–15292, 2021. 1

[20] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 33(12): 2341–2353, 2010. 1, 2, 6

[21] Cong Huang, Jiahao Li, Bin Li, Dong Liu, and Yan Lu. Neural compression-based feature learning for video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5872–5881, 2022. 2, 3

[22] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 106–122, 2018. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[24] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. End-to-end united video dehazing and detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[25] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing (TIP)*, 28(1):492–505, 2018. 2

[26] Boyun Li, Yuanbiao Gou, Jerry Zitao Liu, Hongyuan Zhu, Joey Tianyi Zhou, and Xi Peng. Zero-shot image dehazing. *IEEE Transactions on Image Processing (TIP)*, 29: 8457–8466, 2020. 2

[27] Boyun Li, Yuanbiao Gou, Shuhang Gu, Jerry Zitao Liu, Joey Tianyi Zhou, and Xi Peng. You only look yourself: Unsupervised and untrained single image dehazing neural network. *International Journal of Computer Vision (ICCV)*, 129:1754–1767, 2021. 2

[28] Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. Domain adaptive object detection for autonomous driving under foggy weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 612–622, 2023. 1

[29] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. Single image dehazing via conditional generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8202–8211, 2018. 2

[30] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 1633–1642, 2019. 2

[31] Yunan Li, Qiguang Miao, Wanli Ouyang, Zhenxin Ma, Huijuan Fang, Chao Dong, and Yining Quan. Lap-net: Level-aware progressive network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 3276–3285, 2019. 2

[32] 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 (CVPR)*, pages 4988–4997, 2015. 2

[33] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:378–393, 2022. 3

[34] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2022. 2

[35] 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 (CVPR)*, pages 3233–3242, 2018. 3

[36] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Grid-dehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 7314–7323, 2019. 2

[37] Yang Liu, Jinshan Pan, Jimmy Ren, and Zhixun Su. Learning deep priors for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 2492–2500, 2019. 2

[38] 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 (ACMMM)*, pages 5427–5435, 2022. 2, 6

[39] Earl J McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York*, 1976. 1, 2

[40] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 4

[41] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[42] Ranjan Mondal, Sanchayan Santra, and Bhabatosh Chanda. Image dehazing by joint estimation of transmittance and airlight using bi-directional consistency loss minimized fcn. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 920–928, 2018. 2

[43] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *International journal of computer vision (IJCV)*, 48:233–254, 2002. 1

[44] Yanwei Pang, Jing Nie, Jin Xie, Jungong Han, and Xuelong Li. Bidnet: Binocular image dehazing without explicit disparity estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5931–5940, 2020. 2

[45] Prashant W Patil, Sunil Gupta, Santu Rana, Svetha Venkatesh, and Subrahmanyam Murala. Multi-weather image restoration via domain translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21696–21705, 2023. 2

[46] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 11908–11915, 2020. 2

[47] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12802–12813, 2023. 2

[48] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8160–8168, 2019. 2

[49] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4161–4170, 2017. 3, 5

[50] 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 (TIP)*, 28(4):1895–1908, 2018. 1, 2

[51] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2808–2817, 2020. 2

[52] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Towards domain invariant single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 9657–9665, 2021. 2

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[54] Taeyong Song, Youngjung Kim, Changjae Oh, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Simultaneous deep stereo matching and dehazing with feature attention. *International Journal of Computer Vision (IJCV)*, 128:799–817, 2020. 1, 2

[55] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3360–3369, 2020. 3, 4

[56] 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 (CVPR)*, pages 2353–2363, 2022. 2

[57] 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 (CVPRW)*, pages 0–0, 2019. 3

[58] Qingbo Wu, Jingang Zhang, Wenqi Ren, Wangmeng Zuo, and Xiaochun Cao. Accurate transmission estimation for removing haze and noise from a single image. *IEEE transactions on image processing (TIP)*, 29:2583–2597, 2019. 2

[59] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22282–22291, 2023. 1, 2, 6

[60] Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, and Pheng-Ann Heng. Video dehazing via a multi-range temporal alignment network with physical prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18053–18062, 2023. 2, 3, 6

[61] 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 (CVPR)*, pages 1661–1670, 2019. 3

[62] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2018. 2

[63] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings*

[64] Yijun Yang, Angelica I Aviles-Rivero, Huazhu Fu, Ye Liu, Weiming Wang, and Lei Zhu. Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13200–13210, 2023. 2

[65] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision (ECCV)*, pages 130–145. Springer, 2022. 2

[66] Hu Yu, Jie Huang, Yajing Liu, Qi Zhu, Man Zhou, and Feng Zhao. Source-free domain adaptation for real-world image dehazing. In *Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, pages 6645–6654, 2022. 2

[67] Jiyang Yu, Jingen Liu, Liefeng Bo, and Tao Mei. Memory-augmented non-local attention for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17834–17843, 2022. 4

[68] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3194–3203, 2018. 2

[69] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Joint transmission map estimation and dehazing using deep networks. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(7):1975–1986, 2019. 2

[70] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *European Conference on Computer Vision (ECCV)*, pages 581–596. Springer, 2022. 3, 4

[71] 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 (CVPR)*, pages 9239–9248, 2021. 2, 3

[72] Shiyu Zhao, Lin Zhang, Ying Shen, and Yicong Zhou. Refinednet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing (TIP)*, 30:3391–3404, 2021. 2, 6

[73] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[74] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 2482–2491, 2019. 4

[75] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing (TIP)*, 24(11):3522–3533, 2015. 2