Learning to Restore Hazy Video: A New Real-World Dataset and A New Method

Xinyi Zhang[1*] Hang Dong[2,3*†] Jinshan Pan[4] Chao Zhu[3] Ying Tai[1] Chengjie Wang[1]
Jilin Li[1] Feiyue Huang[1] Fei Wang[3]
[1] Tencent Youtu Lab [2] ByteDance Intelligent Creation Lab
[3] College of Artificial Intelligence, Xi'an Jiaotong University
[4] Nanjing University of Science and Technology

Abstract

Most of the existing deep learning-based dehazing methods are trained and evaluated on the image dehazing datasets, where the dehazed images are generated by only exploiting the information from the corresponding hazy ones. On the other hand, video dehazing algorithms, which can acquire more satisfying dehazing results by exploiting the temporal redundancy from neighborhood hazy frames, receive less attention due to the absence of the video dehazing datasets. Therefore, we propose the first REal-world VIdeo DEhazing (REVIDE) dataset which can be used for the supervised learning of the video dehazing algorithms. By utilizing a well-designed video acquisition system, we can capture paired real-world hazy and haze-free videos that are perfectly aligned by recording the same scene (with or without haze) twice. Considering the challenge of exploiting temporal redundancy among the hazy frames, we also develop a Confidence Guided and Improved Deformable Network (CG-IDN) for video dehazing. The experiments demonstrate that the hazy scenes in the REVIDE dataset are more realistic than the synthetic datasets and the proposed algorithm also performs favorably against state-of-the-art dehazing methods.

1. Introduction

Images and videos captured from the hazy scenes inevitably suffer from limited visibility and low color saturation due to the particles in the haze that will scatter and absorption the light and decrease the albedo of the viewed scene. The goal of the dehazing algorithms is to remove the haze and restore a haze-free scene by given a hazy image or video. This problem has received significant attention since the dehazing algorithm is a necessary pre-processing step for many high-level vision tasks (e.g., scene understanding [31] and detection [18]) applied on the outdoor haze, indoor fire, and smoking scenes.

Recently, the introduction of new techniques from machine learning and deep learning provides a broader perspective for dehazing problem and achieves impressive results. Existing deep learning-based methods [9, 27, 40, 17, 38] are usually trained on the synthetic datasets [19], in which the hazy scene I is formulated by:

$$I(x) = T(x)J(x) + (1 - T(x))A, \tag{1}$$

where J denotes the haze-free scene, A describes the global atmospheric light indicating the intensity of ambient light, T is the transmission map, and x represents the pixel position. However, the scattering atmosphere model in Equ. (1) has shown several limitations: it cannot formulate realistic hazy scenes with active light sources [20], with non-homogeneous haze [5], with dense haze layer [4], and under complex illumination conditions. Therefore, the networks trained on these synthetic datasets often generate unsatisfied results when handling real-world inputs due to the domain shift [34]. Recently, Some realistic image dehazing datasets [6, 3, 4, 5] are introduced to provide benchmarks for training and evaluating real-world dehazing algorithms. Since then, great progress has been made in the study of the real-world image dehazing task [2, 7, 8].

Although significant achievements have been made in single image dehazing task, we believe that video dehazing algorithms can achieve better results by utilizing the temporal redundancy from neighboring frames. However, due to the difficulty of collecting real-world video dehazing datasets, the video dehazing task receives less attention than image dehazing [32]. Although plenty of synthetic hazy videos can be obtained by using Equ. (1) [19], the domain gap between synthetic and real-world hazy videos makes these synthetic datasets low practical value. Therefore, collecting a real-world video dehazing dataset for deep learning-based algorithms is a challenging but valuable work.

In this paper, we build a Consecutive Frames Acquisition System (CFAS), which can be used for collecting paired videos via a controllable robot arm. By utilizing the accurate

*These authors contributed equally to this work.
†Corresponding author.

relocation ability of the robot arm, the system can record the acquisition points of the last collected video and collect another but exactly the same video if the scene does not change. With the newly-designed video acquisition system and professional haze machines, we can collect the pairs of real hazy and corresponding haze-free videos by generating high fidelity haze between the acquisition of the two videos. By collecting real hazy and corresponding haze-free videos on various indoor scenes, we contribute the REal-world VIdeo DEhazing (REVIDE) Dataset, the first video dehazing dataset for supervised learning. Both subjective and objective experiments indicate that the REVIDE dataset contains more realistic hazy frames than the synthetic one, which can help the training and evaluating processes of real-world video dehazing algorithms.

Since the haze spreads over the whole scenes and the density of the haze may change across the neighboring frames of a video, temporal alignment and exploiting temporal redundancy are challenging in real-world video dehazing algorithms. In this paper, we present a Confidence Guided and Improved Deformable Network (CG-IDN) for video dehazing. We show that a confidence guided pre-dehazing module and the cost volume [36] can benefit the deformable alignment module by improving the accuracy of the estimated offsets. Moreover, the confidence map can also be used as guidance for multi-feature fusion. Extensive evaluations demonstrate that the proposed algorithm performs favorably against state-of-the-art video and image dehazing methods.

The contributions of this work are summarized as follows:

- In this paper, we collect a real-world video dehazing dataset containing pairs of real hazy and corresponding haze-free videos. To the best of our knowledge, the proposed dataset is the first real-world video dehazing dataset for supervised learning.
- We conduct extensive subjective and objective experiments to demonstrate that the collected hazy scenes in the proposed dataset are more realistic than those of synthetic datasets, which provides a valuable benchmark for training and evaluating real-world video dehazing algorithms.
- We propose a Confidence Guided and Improved Deformable Network (CG-IDN) for video dehazing and validate its effectiveness in real-world video dehazing tasks.

2. Related Work

Dehazing datasets. Recently, deep learning-based approaches have been applied to solve the dehazing problems, which require large-scale dehazing datasets for training and evaluating. However, collecting pairs of the real hazy and corresponding haze-free images is a burdensome work due to the strict constraints on the static state of the illumination condition and viewed scene. To provide enough training data for deep dehazing networks, several large-scale synthetic image dehazing datasets are proposed [19, 1, 27, 17, 32] by utilizing the images and the depth maps of the Middleburry [33], NYU-Depth V2 [35], and Cityscapes [10] datasets. All of these datasets are using Koschmieder's light propagation model [30] to generate synthetic hazy images from haze-free images and the corresponding depth maps. Due to the large domain gap between the synthetic haze and real-world haze, the models trained on these synthetic image dehazing datasets often generate unsatisfied results when handling the real-world hazy scenes.

To reduce the domain gap, several real-world image dehazing datasets [3, 6, 4, 5] are proposed. All of these real-world image datasets use a professional haze machine to imitate with high fidelity real hazy conditions and collect pairs of real-world hazy and corresponding haze-free images (ground truth). However, the acquisition systems applied in these datasets are designed for collecting paired images in a fixed location, which is not suitable for collecting real-world video dehazing dataset.

Compared with image dehazing datasets, the video dehazing dataset is rare. Ren et al. [28] generate a synthetic video dehazing dataset by using the video clips and corresponding depth maps from NYU Depth. Unfortunately, there is still no real-world video dehazing dataset due to the lack of suitable acquisition systems. Since the absence of real-world training data becomes a major obstacle for video dehazing task [32], it is valuable work to collect a large-scale real-world video dehazing dataset.

Video dehazing algorithms. Compared with single image dehazing algorithms, video dehazing algorithms [15, 41, 22] try to generate more accurate dehazed results by taking advantage of the temporal redundancy from neighboring frames. Zhang et al. [41] first propose an algorithm to dehaze the videos frame by frame, and then use the optical flow to improve the temporal coherence of the neighboring frames based on Markov Random Field (MRF). In [22], Li et al. propose an algorithm to jointly estimate scene depth and recover the clear latent image from a foggy video sequence. Recently, deep learning-based methods achieve promising results on many restoration tasks where the large-scale training datasets are available. Based on this, Ren et al. [28] propose a synthetic video dehazing dataset and develop a deep learning solution to accumulate information across frames for transmission estimation. However, their network performs not well in the real-world hazy scenes where the transmission maps are more complex due to the non-homogeneous haze. Wang et al. [37] propose a video restoration network enhanced by the deformable convolutional networks, which achieves superior performance against state-of-the-art methods on several video restoration tasks. However, we found their deformable alignment module suffers from unstable training and fails to estimate large offset when handling frames with large resolution.

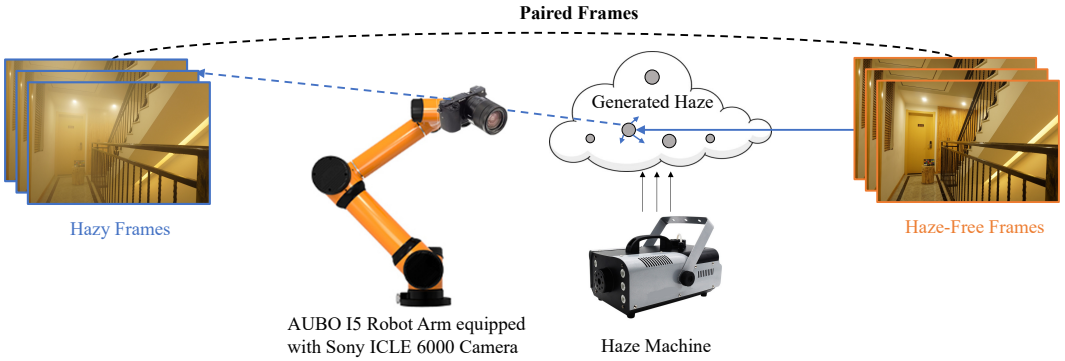

Figure 1. **Consecutive Frames Acquisition System (CFAS).** The system consists of a controllable robot arm (AUBO I5), a Sony ICLE 6000 camera, and two haze machines. By utilizing the accurate relocation ability of the robot arm, we can capture the pairs of hazy and corresponding haze-free videos in the same scene.

3. Real-World Video Dehazing Dataset

To collect a real-world video dehazing dataset for supervised learning, we propose an acquisition system to capture the pairs of hazy and corresponding haze-free videos in the same scene. The detailed system compositions and dataset collection process are presented in the following sections.

3.1. Acquisition system

In order to capture the real-world hazy videos and their corresponding haze-free (ground truth) videos, we design a Consecutive Frames Acquisition System (CFAS). As shown in Fig. 1, the proposed system consists of a controllable robot arm (AUBO I5), a Sony ICLE 6000 camera, and two haze machines. The camera is mounted at the end of the robot arm by a customized mounting rack and remote-controlled by a laptop. Since the AUBO robot arm can repeatedly reach the same location with an accuracy of $2mm$, we can repeat the trajectory of the last collected video to collect the same video [1]. Besides, two kinds of haze machines (DJPOWER E-1500 and DJS-900W) are used to produce high fidelity real hazy conditions.

3.2. Data Collection

Scene layout and system settings. To guarantee the generality of our dataset, all scenes in our dataset are carefully selected. To be specific, scenes with rich and colorful textures are preferred since the primary purpose of dehazing method is to recover high-frequency details from the hazy inputs. As shown in Fig. 2, the selected scenes can be grouped into four styles: the Eastern style, the Western style, the Laboratory style, and the Corridor style. The whole dataset contains 47 different scenes, which contain different layouts, illumination conditions, and density of the haze.

[1]According to the relocation accuracy of the AUBO robot and the parameters of the camera, we can conclude that the shift on the pixel whose depth is larger than 2.8m will be less than 1 pixel even when the relocation error occurs.

Figure 2. **Examples of different styles of scenes in the REVIDE Dataset**.

Table 1. **Number of pairs for training and testing sets in the REVIDE dataset.**

Scenes Style	Esatern style	Western style	Laboratory style	Corridor style
Tran Set	382	575	498	243
Test Set	57	137	54	36

Since capturing a video dehazing dataset needs more strict constraints on the static state of the viewed scenes, each scene should be carefully set. The windows and doors must be closed to let the scene keep isolated, and some lightweight objects are removed as they might be moved by the airflow of the haze machines. Once the scene layout is done, the objects and the illumination conditions of the viewed scene should be static during the acquisition process. Since the static state of outdoor scenes cannot be guaranteed, we only collect indoor scenes for the proposed dataset at this time. More details about the optical parameter of the camera can be found in the supplementary material.

Acquisition process. After finishing all the preparatory works, we run a multi-threading program to move the robot arm and take remote control of the camera to capture the sharp frames at the acquisition points of a planned trajectory. The timeline of the acquisition process can be summarized as Fig. 3. The program will take 0.55 seconds to initial the robot arm and the video collection process is activated 1 second after the arm starts to move. The number of the acquisition points ranges from 50 to 100, depending on the trajectories, which means one captured video contains 50-100 consecutive frames. The whole haze-free video acquisition

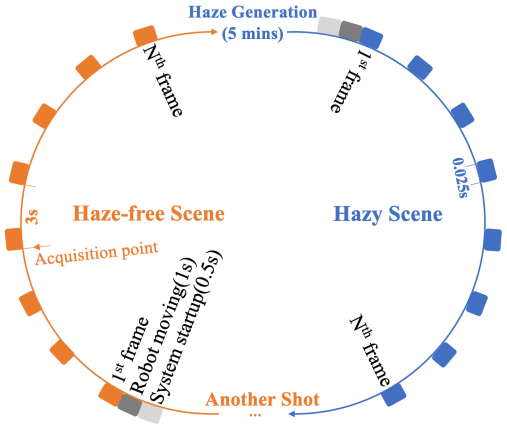

Figure 3. **Timeline of the acquisition process.** To capture the pairs of hazy and corresponding haze-free videos, the camera will move along the same trajectory twice and collect the hazy and haze-free frames.

process usually takes 3-5 minutes, and then the robot arm is re-initialized to the starting point for the next acquisition.

Before the second acquisition, the haze machines are activated for 1-2 minutes, and we shall wait approximately 3 minutes to let the haze fully spread around the room. Then, the corresponding hazy frames are collected by running the multi-process program for the second time. Relying on the relocation ability of the robot arm, every acquisition point of the first acquisition process is recorded and it can be precisely relocated when collecting the corresponding hazy video frames. Thus, the alignment between the hazy video and corresponding haze-free video can be guaranteed. It takes us one month to collect 48 video pairs from scenes with four different styles[2].

After checking the collected videos throughout, the bad frames which contain geometric misalignment, aberration, undesired blur, and over-exposure are discarded. For standardization, all the collected video frames are cropped to 2708×1800. The numbers of pairs for training and testing sets are listed in Tab. 1.

4. Confidence Guided and Improved Deformable Network

In this section, we describe the architecture design, training loss functions, and implementation details of the proposed CG-CDN for video dehazing.

4.1. Network Architecture

Given $2N+1$ hazy frames $I_{[t-N:t+N]}$ as the input, our goal is to recover a haze-free result $\hat{J}_t$ of the reference frame (i.e., middle frame). As illustrated in Fig. 4, the proposed model consists of four modules:

- A Confidence Guided Pre-Dehazing (CGPD) module for pre-processing the hazy frames and estimating the confidence maps of the reference frame C_t.
- A Improved Deformable Alignment (IDA) module to align the enhanced features $F_{[t-N:t+N]}$ from the CGPD module by taking the partial cost volumes as guidances.
- A Multi-Feature Fuison (MFF) module to fuse the aligned features $F^{Align}_{[t-N:t+N]}$ from the IDA module by leveraging the confidence map C_t.
- A restoration module to reconstruct haze-free result of the reference frame $\hat{J}_t$ from the fused features F^{Fused}_t.

Confidence guided pre-dehazing module. Since the density of the haze may change across the neighboring frames, it is necessary to pre-process the hazy inputs to improve the performance of the following alignment module [37]. As shown in Fig. 4, the Confidence Guided Pre-Dehazing (CGPD) module is built with three Confidence Blocks (CBs). For each confidence block i, the enhanced features from the last confidence block $F^{i-1}_{[t-N:t+N]}$ are enhanced under the guidance of the confidence maps from the last confidence block $C^{i-1}_{[t-N:t+N]}$. Then, the pre-dehazing results $\hat{J}^i_{[t-N:t+N]}$ and confidence maps $C^i_{[t-N:t+N]}$ of block i are separately generated by two output head, the reconstruction head, and confidence head. The enhnaced features $F^i_{[t-N:t+N]}$ and confidence maps $C^i_{[t-N:t+N]}$ are fed into the next confidence block for further enhancement. The parameters are shared when handling different frames and the input confidence maps of the first confidence block are set as 0.

For simplicity, we define the output enhanced features and confidence maps of the last confidence block as $F_{[t-N:t+N]}$ and $C_{[t-N:t+N]}$. The enhanced features $F_{[t-N:t+N]}$ are send to the improved deformable alignment module for feature-level alignment. The confidence map of the reference frame C_t, which is used to describe the fidelity of each pixel of the pre-dehazing result, is sent to the multi-feature fusion module.

Improved deformable alignment module. To address the unstable training issue of the PCD deformable alignment module [37], we proposed an Improved Deformable Alignment (IDA) module by introducing the partial cost volume [36] to each level of the PCD. As shown in Fig. 5, the IDA module introduces a correlation layer and three scale-sampling layers (in the dashed box) to each level of the PCD module. More specifically, the partial cost volume between the enhanced features of the neighboring frame F_{t+n} and the reference frame F_t is calculated by the correlation layer [36] and then it is concatenated with the F_t to estimate the offsets for the deformable convolution (DConv). It is also noted that the offsets between two adjacent frames in our video dehazing dataset may be larger than 100 pixels, and it is computationally expensive to compute the partial cost volume with a range larger than 100 pixels. To maintain accuracy while reducing the computational cost, we first

[2]To enrich our dataset, we will add 6 outdoor scenes (4 for training and 2 for evaluation) in the released version of the REVIDE dataset.

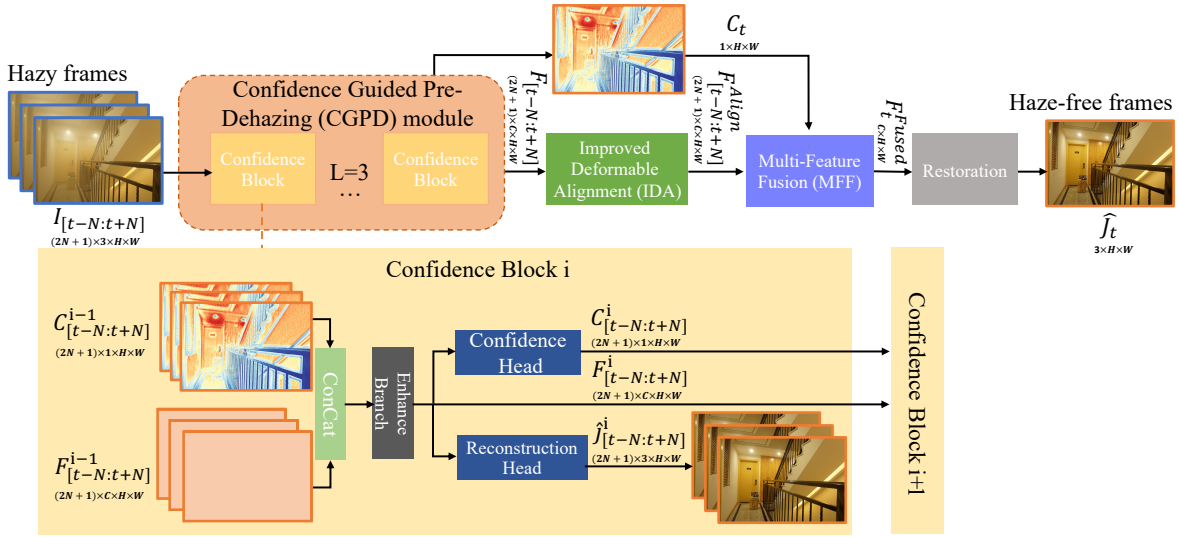

Figure 4. **Architecture of the proposed Confidence Guided and Improved Deformable Network (CG-IDN) for video dehazing.**

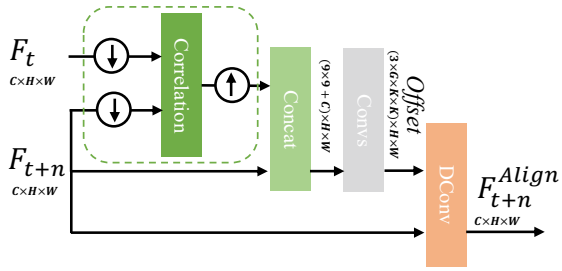

Figure 5. **Details of the Improved Deformable Alignment (IDA) module.**

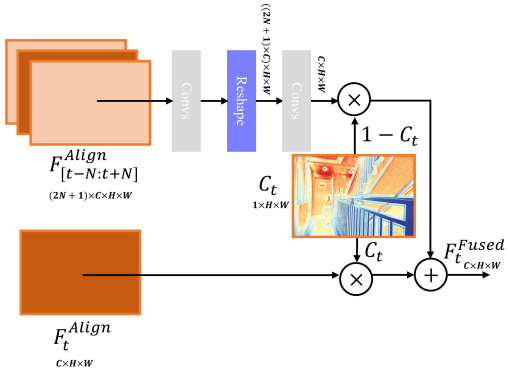

Figure 6. **Details of the Multi-Feature Fuison (MFF) module.**

downsample the F_{t+n} and F_t to 1/16 of the original resolution with bilinear upsampling layer and calculate the partial cost volume with a range of 9 pixels. Then, the partial cost is upsampled to the resolution of F_t via the nearest upsampling layer.

Since the cost volume is a more discriminative representation of the dense correspondences, we believe that the IDA module can obtain a more robust aligned feature F_{t+n}^{Align}.

Multi-feature fuison module. To fully exploit the temporal redundancy, a Multi-Feature Fusion (MFF) module is proposed to fuse the aligned neighboring features $F_{t-N,t+N}^{Align}$ from IDA module. As shown in Fig. 6, the fusion process can be formulated as:

$$F_t^{Fused} = F_t^{Align} \times C_t + \phi_\theta(F_{[t-N,t+N]}^{Align}) \times (1 - C_t), \quad (2)$$

where the F_t^{Fused} is the fused output of the MFF module, F_t^{Align} and C_t is the aligned feature and confidence map of the reference frame, and ϕ_θ denotes the operations $(Conv - Reshape - Conv)$ to extract useful temporal redundancy. By introducing the confidence map as guidance, the MFF module can reserve the high fidelity features of the reference frame and enhance the uncertain features by fusing the temporal redundancy from neighboring features.

Restoration module. In the final stage of the proposed CG-IDN, the fused features F_t^{Fused} are used to reconstruct the haze-free result $\hat{J}_t$ via a restoration module. Since the resolution of videos in the proposed REVIDE dataset is relatively large (2708×1800), the restoration module with large receptive field and computational efficiency is preferred. Therefore, we choose the MSBDN [11], a state-of-the-art image dehazing method based on the U-Net [29] architecture, as our restoration module. To reduce the parameters, the DFF modules in MSBDN are removed. Since most operations of MSBDN are performed at 1/16 resolution, the haze-free results $\hat{J}_t$ can be reconstructed with a large receptive field and low computational cost in both training and inference phases.

4.2. Loss Function

The proposed network generates three kinds of outputs: the pre-dehazing results and confidence maps of three con-

fidence blocks ($\hat{J}^i_{[t-N:t+N]}$ and $C^i_{[t-N:t+N]}, i \in \{1,2,3\}$) and the final haze-free result $\hat{J}_t$. Since the REVIDE dataset provides haze-free ground truth for each hazy frame, we can train the network in a supervised manner. Thus, the overall objective of the CG-IDN can be denoted as:

$$L = L_d + \lambda_p L_p + \lambda_c L_c, \tag{3}$$

where L_d and L_p denote the L1 loss and perceptual loss between the haze-free results $\hat{J}_t$ and ground truth J_t, L_c is loss function for confidence blocks, and λ_p and λ_c is the weight to balance these three loss terms.

According to [39], the L_c for three confidence blocks can be defined as:

$$L_c = \frac{1}{3(2N+1)} \sum_{i=1}^{3} \sum_{n=-N}^{N} C^i_{t+n} (\hat{J}^i_{t+n} - J_{t+n})^2 - \lambda_r \log C^i_{t+n}, \tag{4}$$

where i is the ordering of the confidence block, J_{t+n} is the haze-free ground truths of all the input frames, and λ_r is a weight factor. By minimizing L_c, the confidence blocks are encouraged to produce pre-dehazing results $\hat{J}^i_{[t-N:t+N]}$ that are close to the haze-free ground-truth $J^i_{[t-N:t+N]}$ and give high confidence scores to the high-fidelity dehazing results. During the training process, we empirically set λ_p to 0.5, λ_c to 0.5, and λ_r to 0.01.

4.3. Implementation

The enhanced branch in confidence block is built with three residual blocks [37]. Both the confidence and dehazing heads consist of two convolutional layers, except that the confidence head ends with a sigmoid activation layer. To match the channel size configuration of MSBDN, the channel size (C) of the CGPD, IDA module, and MFF modules are set to 16. The Leaky Rectified Linear Unit (LReLU) with a negative slope of 0.1 is used after each convolutional and deformable convolutional layers. When trained on the RIVIDE dataset, the network takes three consecutive frames (i.e., N=1) as inputs due to the large displacement between the adjacent frame.More implementation details can be found in the supplementary material.

To augment the training data, we resize the hazy inputs and ground truths with three random scales within a scale of 0.5 and 1.0. We crop the hazy inputs and ground truths to patches with a size of 384×384 and augment them with random horizontal flips and 90° rotations. The entire training process contains 100 epochs optimized by the ADAM solver [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a batch size of 4. The initial learning rate is set as 10^{-4} with a decay rate of 0.1 after every 45 epochs. The proposed network is implemented with our the PyTorch framework [25] and is trained and evaluated on an NVIDIA TITAN RTX GPU. The REVIDE dataset and the code are available at http://xinyizhang.tech/revide.

5. Experiments

In this section, we conduct objective and subjective experiments to validate the fidelity of the hazy scenes in the REVIDE dataset. In addition, for evaluating the proposed dataset and algorithms, we train the proposed CG-IDN and other state-of-the-art dehazing approaches on the proposed REVIDE dataset to present the quantitative and qualitative comparisons.

5.1. REVIDE vs. Synthetic Video Dehazing Dataset

For fair comparisons, we generate a synthetic video dehazing dataset using the same scenes of the REVIDE dataset. Specifically, we use [21] to estimate the depth maps $d(x)$ of the haze-free frames in the REVIDE dataset and generate the transmission maps by sampling β between [1.6, 3.4]. Then, the corresponding synthetic hazy frames are generated with the same configurations in the RESIDE dataset [19]. This synthetic video dehazing dataset is referred to as REVIDE-SYN.

Objective evaluation. To demonstrate that the collected hazy scenes in REVIDE dataset are more realistic and can better represent the real-world hazy scenes than the synthetic datasets, we train a binary classification network, based on a pre-trained DenseNet-121 [14], to distinguish the real-world hazy scenes from synthetic hazy scenes. The training set consists of 1384 real-world outdoor hazy scenes from the Unannotated Real-world Hazy Images (URHI) dataset in the RESIDE dataset [19] and 1950 synthetic indoor scenes from the OTS set of RESIDE dataset [19] and the Binary Cross Entropy (BCE) loss is adopted to optimize the network parameters. During the testing stage, we use the trained model to classify 1150 real-world outdoor hazy scenes (referred to as ROS set) from the internet, 1150 collected indoor hazy scenes (referred to as CIS set) from the REVIDE dataset, and 1150 synthetic indoor hazy scenes (referred to as SIS set) from the REVIDE-SYN dataset. To avoid the interference of different scenes, the CIS and SIS contain the same scenes and similar haze density. The average ratios and probability that the scenes of each test set are classified as real-world hazy scenes are presented in Tab. 2. Although the training dataset does not contain any real-world indoor hazy images, most collected scenes of the CIS set (80.1%) are classified as real-world scenes by the trained classification network. On the other hand, only 24.2% of the synthetic scenes in the CIS set are classified as real-world scenes. Therefore, the classification results demonstrate that the proposed REVIDE dataset is more realistic than the synthetic datasets and can generalize well to the real-world hazy scenes.

Subjective evaluation. We also performed a user study to compare the fidelity of the hazy scenes in the ROS, CIS, and SIS sets with human perception. We sample one hazy scene from ROS, CIS, and SIS respectively, and combine them into a triplet with a random order. Finally, 500 triplets are obtained and evenly divided into 10 groups. For each

Table 2. **Classification results on ROS, CIS, and SIS.** According to the outputs of the trained binary classification network, this table presents the average ratios and probability that the scenes of each test set are classified as real-world hazy scenes.

	Avg. Ratios	Avg. Probability
ROS	**96.9%**	0.921
CIS	**80.1%**	0.759
SIS	24.2%	0.297

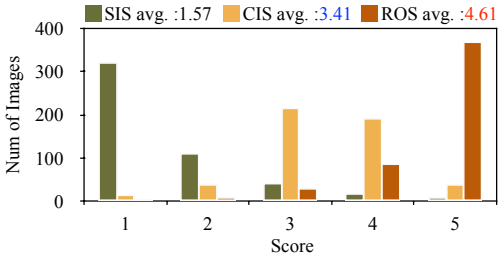

Figure 7. **Average and distribution of the mean opinion scores**

group, five experienced researcher in computer vision, aged between 22 and 56, are requested to give scores to the hazy scenes in the triplets according to the fidelity of the haze. The score is within a range of 1 to 5, from low fidelity to high fidelity. As shown in the top of Fig. 7, the average score of CIS is 3.408 and is closer to the score of ROS (4.61) than the score of SIS (1.57), which demonstrates that the collected hazy scenes can successfully imitate the real-world hazy conditions. The bars with different colors in Fig. 7 display the distributions of the scores for each set. As the chart shows, most CIS scenes obtain positive scores (3-5), for the reason that the collected hazy scenes are more natural when the active light source, non-homogeneous haze, dense haze layer, and complex illumination conditions exist. On the other hand, the scores of SIS is concentrated in the negative scores (1-2), because that the synthetic hazy scenes always suffer from low color saturation, unnaturally haze distribution, and inconsistent color temperature between scene and synthetic haze. Some typical scenes from the REVIDE and REVIDE-SYN will be presented in the supplementary material for better illustration. According to the objective and subjective evaluations, the proposed REVIDE dataset contains realistic hazy scenes and can be used to evaluate whether a dehazing algorithm have a generalization ability to handle the real hazy scenes.

5.2. Performance Evaluation

To demonstrate the advantages of the proposed REVIDE dataset and the effectiveness of the proposed method, we evaluate the proposed CG-IDN on the REVIDE against several competitive image dehazing methods (DCP [12], KDDN [13], GDN [23], DuRN [24], FFA [26], and MSBDN [11]) and video dehaing methods (VDN [28] and EDVR [37]). Except for the DCP, we re-train all the deep learning-based methods on the REVIDE and REVIDE-SYN dataset separately and then evaluate all the models on the test set of the REVIDE dataset.

Evaluation on the REVIDE dataset. The first row in Tab. 3 shows the quantitative results of models trained on the synthetic dataset, REVIDE-SYN. Although these models have achieved favorable results on the test set of the REVIDE-SYN dataset, they cannot generalize well to the realistic hazy videos in REVIDE dataset. Most methods only obtain unsatisfactory results that are even worse than the hazy inputs and only the VDN [28] acquires valid results by estimating the transmission map of the reference frame instead of the dehazed frame directly. These quantitative results indicate that the real-world hazy scenes are difficult to simulate and the models train on synthetic dataset does not perform well on the realistic datasets due to domain gap.

The second row in Tab. 3 shows the quantitative results of models trained on the REVIDE dataset. The DCP does not perform well since the dark channel prior does not hold for most indoor scenes. Since the REVIDE dataset can provide a high fidelity training set, most deep learning-based methods trained on it achieve valid results compared with the hazy inputs. Among these image dehazing methods [26, 11, 13, 23, 24], the MSBND [11] achieves the most competitive results due to its U-Net architecture is more suitable for handling inputs with large resolution (2708×1800). The MSBND even outperforms the video dehazing methods [28, 37], which indicates the importance of the network architecture. The VDN does not perform well because the transmission maps in the real-world hazy videos are more complex than those in the synthetic videos. The unified framework for video restoration, EDVR, suffers from unstable training and fails to estimate large offset when trained on the REVIDE dataset. Benefitting from the confidence-guided strategy and the improved deformable alignment module, the proposed CG-IDN can effectively align the neighboring features and fuse the aligned features. Therefore, The CG-IDN obtains the best performance on the REVIDE dataset.

Fig. 8 shows two visual examples from the test set of the REVIDE dataset. The DCP algorithm only obtains meaningless results with significant color distortions due to the dark channel prior does not hold for most indoor scenes. The dehazed images by most deep learning frameworks [23, 28, 37] still contain significant remaining haze and artifacts. Although the MSBDN method successfully removes most haze, its results suffer from inconsistent color temperature with the haze-free ground truths. In contrast, our algorithm obtains clean results without color distortions due to successfully exploiting the temporal redundancy from neighboring frames. More qualitative results on the REVIDE dataset and real-world videos can be found in the supplementary material.

Ablation study. We perform ablation studies to analyze the importance of each key component in CG-IDN. All the methods mentioned below are trained on the REVIDE dataset using the same setting.

Table 3. **Quantitative evaluations on the REVIDE dehazing datasets.** The first and second rows show the quantitative results of models trained on the training set of the REVIDE-SYN dataset and REVIDE dataset, respectively. All the models are trained on the test set of the REVIDE dataset. **Red texts** and blue texts indicate the best and the second-best performance respectively.

Methods		DCP [12]	GDN [23]	DuRN [24]	KDDN [13]	MSBDN [11]	FFA [26]	VDN [28]	EDVR [37]	CG-IDN (Ours)
Trained on Syn.	PSNR	11.03	14.02	14.54	14.40	13.41	11.17	**17.54**	14.95	15.02
	SSIM	0.7285	0.7437	0.7776	0.7542	0.7211	0.5685	**0.8269**	0.7908	0.7992
Trained on Real.	PSNR	11.03	19.69	18.51	16.32	22.01	16.65	16.64	21.22	**23.21**
	SSIM	0.7285	0.8545	0.8272	0.7731	0.8759	0.8133	0.8133	0.8707	**0.8836**

(a) Hazy input (b) DCP [12] (c) GDN [23] (d) MSBDN [11] (e) VDN [28] (f) EDVR [37] (g) Ours (h) Ground-truth

Figure 8. **Visual results on the REVIDE dataset.** DCP [12] does not perform well since the dark channel prior does not hold for most indoor scenes. The results in (c)-(g) contain some color distortions and haze residual, while the dehazed image in (h) by our method is much clearer. Best viewed on a high-resolution display. More video results can be found in the supplementary material.

Table 4. **Analysis on each component of the proposed CG-IDN. Red texts** indicate the best performance of each part.

Methods	Baseline-F1	Baseline-F3	Baseline-PWC	Baseline-DA	Baseline-IDA	Baseline-PD-IDA	CG-IDN
Multi-frame input		✓	✓	✓	✓	✓	✓
Optical-fow based			✓				
Deformable alignment				✓	✓	✓	✓
Cost volume					✓	✓	✓
CGPD						✓	✓
MFF							✓
Paramters	21.6M	21.6M	31M	21.9M	21.9M	22.9M	23M
PSNR	21.81	21.85	22.10	22.42	22.86	23.08	**23.21**

We first train the restoration module of the CG-IDN as the baseline model (referred to as Baseline-F1 and Baseline-F3 according to the number of the input frames). As shown in the Tab. 4, Baseline-F3 achieves comparable results with Baseline-F1, which demonstrates that the network cannot exploit the temporal information from neighboring frames without an alignment module. To align the neighboring frames, the PWC-Net [36] and warping layer are introduced to the Baseline-F3, which is referred to as the Baseline-PWC. We also introduce the deformable alignment module in EDVR [37] to perform frame alignment at feature level (Baseline-DA). The Baseline-DA outperforms Baseline-F3 and Baseline-PWC by a margin of 0.57 dB and 0.32 dB respectively, which demonstrates the effectiveness of the deformable alignment. However, we find the deformable alignment is not optimal and it fails to estimate the offset when trained on the REVIDE dataset (see Section C in the supplementary). Therefore, we improve the deformable alignment by introducing partial cost volume as the guidance for the offset estimation (Baseline-IDA). Compared with the Baseline-DA, the Baseline-IDA achieves 0.44 dB gain by only slightly increasing the computational cost. Finally, the proposed network is constructed from Baseline-IDA by successively introducing the CGPD (Baseline-PD-IDA) and MMF (CG-IDN) module. Both the Baseline-PD-IDA and CG-IDN achieve 0.22 dB and 0.35 dB performance gain over the Baseline-IDA, which demonstrates that the CGPD and MMF can boost the dehazed results via facilitating the temporal alignment and feature fusion processes.

6. Conclusions

We have presented the first REal-world VIdeo DEhazing (REVIDE) dataset collected by a well-designed Consecutive Frames Acquisition System (CFAS). The REVIDE dataset contains pairs of real hazy and corresponding haze-free videos, which can be used for training and evaluating the video dehazing algorithms. Based on the REVIDE dataset, we also develop a Confidence Guided and Improved Deformable Network (CG-IDN) by utilizing the confidence maps and cost volume to boost the dehazing performance. Both subjective and objective evaluations show that the REVIDE dataset contains more realistic hazy scenes than the synthetic datasets and the models trained on it can generalize well to real-world hazy scenes. Extensive experiments demonstrate that the proposed CG-IDN performs favorably against state-of-the-art methods on the REVIDE dataset.

Acknowledgements. H. Dong, C. Zhu, and F. Wang are supported in part by National Major Science and Technology Projects of China (No. 2019ZX01008101) and the Fundamental Research Funds for the Central Universities (FRFCU). J. Pan is supported in part by NSFC (Nos. 61872421, 61922043), the FRFCU (No. 30920041109), and the National Key Research and Development Program of China (No. 2018AAA0102001).

References

[1] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer. D-hazy: A dataset to evaluate quantitatively dehazing algorithms. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2226–2230, 2016. 2

[2] Cosmin Ancuti, Codruta O Ancuti, and Radu Timofte. Ntire 2018 challenge on image dehazing: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 891–901, 2018. 1

[3] Cosmin Ancuti, Codruta O Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: a dehazing benchmark with real hazy and haze-free indoor images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 620–631. Springer, 2018. 1, 2

[4] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1014–1018. IEEE, 2019. 1, 2

[5] 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*, pages 444–445, 2020. 1, 2

[6] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 754–762, 2018. 1, 2

[7] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, Luc Van Gool, Lei Zhang, and Ming-Hsuan Yang. Ntire 2019 image dehazing challenge report. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[8] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 490–491, 2020. 1

[9] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 1

[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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 2

[11] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2157–2167, 2020. 5, 7, 8

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. 7, 8

[13] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462–3471, 2020. 7, 8

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 6

[15] Jin-Hwan Kim, Won-Dong Jang, Yongsup Park, Dong-Hahk Lee, Jae-Young Sim, and Chang-Su Kim. Temporally x real-time video dehazing. In *2012 19th IEEE International Conference on Image Processing*, pages 969–972. IEEE, 2012. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[17] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *IEEE International Conference on Computer Vision*, pages 4770–4778, 2017. 1, 2

[18] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. End-to-end united video dehazing and detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 7016–7023, 2018. 1

[19] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Reside: A benchmark for single image dehazing. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 1, 2, 6

[20] Yu Li, Robby T Tan, and Michael S Brown. Nighttime haze removal with glow and multiple light colors. In *IEEE International Conference on Computer Vision*, pages 226–234, 2015. 1

[21] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 6

[22] Zhuwen Li, Ping Tan, Robby T. Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4988–4997, 2015. 2

[23] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *IEEE International Conference on Computer Vision*, pages 7314–7323, 2019. 7, 8

[24] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 7, 8

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6

[26] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for

single image dehazing. In *AAAI*, pages 11908–11915, 2020. 7, 8

[27] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European Conference on Computer Vision*, pages 154–169, 2016. 1, 2

[28] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE Transactions on Image Processing*, 28(4):1895–1908, 2018. 2, 7, 8

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015. 5

[30] G. C S. Beiträge zur physik der freien atmosphäre. *Nature*, 72(1855):53–53, 1905. 2

[31] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, pages 1–20, 2018. 1

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 1, 2

[33] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 195–202, 2003. 2

[34] 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*, pages 2808–2817, 2020. 1

[35] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012. 2

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2, 4, 8

[37] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4, 6, 7, 8

[38] Dong Yang and Jian Sun. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In *European Conference on Computer Vision*, pages 702–717, 2018. 1

[39] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8405–8414, 2019. 6

[40] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2018. 1

[41] Jiawan Zhang, Liang Li, Yi Zhang, Guoqiang Yang, Xiaochun Cao, and Jizhou Sun. Video dehazing with spatial and temporal coherence. *The Visual Computer*, 27(6-8):749–757, 2011. 2