Deep Video Dehazing With Semantic Segmentation

Wenqi Ren, *Member, IEEE*, Jingang Zhang, Xiangyu Xu, Lin Ma, *Member, IEEE*, Xiaochun Cao, *Senior Member, IEEE*, Gaofeng Meng, *Senior Member, IEEE*, and Wei Liu, *Member, IEEE*

***Abstract*— Recent research have shown the potential of using convolutional neural networks (CNNs) to accomplish single image dehazing. In this paper, we take one step further to explore the possibility of exploiting a network to perform haze removal for videos. Unlike single image dehazing, video-based approaches can take advantage of the abundant information that exists across neighboring frames. In this paper, assuming that a scene point yields highly correlated transmission values between adjacent video frames, we develop a deep learning solution for video dehazing, where a CNN is trained end-to-end to learn how to accumulate information across frames for transmission estimation. The estimated transmission map is subsequently used to recover a haze-free frame via atmospheric scattering model. In addition, as the semantic information of a scene provides a strong prior for image restoration, we propose to incorporate global semantic priors as input to regularize the transmission maps so that the estimated maps can be smooth in the regions of the same object and only discontinuous across the boundaries of different objects. To train this network, we generate a dataset consisted of synthetic hazy and haze-free videos for supervision based on the NYU depth dataset. We show that the features learned from this dataset are capable of removing haze that arises in outdoor scenes in a wide range of videos. Extensive experiments demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods on both synthetic and real-world videos.**

***Index Terms*— Video dehazing, defogging, transmission map, convolutional neural network.**

Manuscript received January 11, 2018; revised June 26, 2018 and September 14, 2018; accepted October 6, 2018. Date of publication October 15, 2018; date of current version December 12, 2018. This work was supported in part by the National Key R&D Program of China under Grant 2016YFB0800603, in part by the National Natural Science Foundation of China under Grant 61802403, Grant U1605252, Grant U1736219, and Grant 61733007, in part by the Key Program of the Chinese Academy of Sciences under Grant QYZDB-SSW-JSC003. The work of J. Zhang was supported in part by the Joint Foundation Program of the Chinese Academy of Sciences for equipment pre-feasibility study under Grant 6141A01011601, in part by the the National Natural Science Foundation of China under Grant 61775219 and Grant 61640422, and in part by the Equipment Research Program of the Chinese Academy of Sciences under Grant Y70X25A1HY. The work of W. Ren was supported in part by the Open Projects Program of National Laboratory of Pattern Recognition and in part by the CCF-Tencent Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dong Xu. *(Wenqi Ren, Jingang Zhang, and Xiangyu Xu contributed equally to this work.) (Corresponding authors: Xiaochun Cao; Gaofeng Meng.)*

W. Ren and X. Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: renwenqi@iie.ac.cn; caoxiaochun@iie.ac.cn).

J. Zhang is with the Medical School, University of Chinese Academy of Sciences, Beijing 100080, China, (e-mail: zhangjg@ucas.ac.cn).

X. Xu is with SenseTime Research, Beijing 100084, China (e-mail: xuxiangyu2014@gmail.com).

G. Meng is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gfmeng@nlpr.ia.ac.cn).

L. Ma and W. Liu are with the Tencent AI Laboratory, Shenzhen 518057, China (e-mail: forest.linma@gmail.com; wliu.cu@gmail.com).

Digital Object Identifier 10.1109/TIP.2018.2876178

I. Introduction

OUTDOOR images and videos often suffer from limited visibility due to haze, fog, smoke, and other small particles in the air that scatter the light in the atmosphere [1]–[5]. Haze has two effects on the captured videos: it attenuates the signal of the viewed scene, and it introduces an additive component to the image, termed the atmospheric light (the color of a scene point at infinity). The image degradation caused by haze increases with the distance from the camera, since the scene radiance decreases and the atmospheric light magnitude increases. Thus, a single hazy image or frame can be modeled as a per-pixel combination of a haze-free image, scene transmission map and the global atmospheric light as follow [6], [7],

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

where $\boldsymbol{I}(x)$ and $\boldsymbol{J}(x)$ are the observed hazy image and the clear scene radiance, $\boldsymbol{A}$ is the global atmospheric light, and $t(x)$ is the scene transmission describing the portion of light that is not scattered and reaches the camera sensors.

Our goal is to recover haze-free frames and corresponding transmission maps. This is an ill-posed problem since there are at least three unknowns per pixel, with inherent ambiguity between haze and object radiance [8]. To handle this highly under-constrained problem, numerous haze removal methods have been proposed [9]–[15] in recent years with significant advancements. Some previous works use additional information such as more images, while others assumed an image prior to solve the problem from a single image [16]–[19].

The most successful video dehazing approaches use information from neighboring frames to estimate transmission maps from the input video [20], taking advantage of a hazy video is temporally coherent and thus the transmissions of an object are similar between adjacent image frames. Based on this assumption, one can design the temporal coherence constraint and add it to the loss costs. Then, the optimal transmissions for each frame can be obtained by minimizing the overall cost [20]. One of the main challenges associated with aggregating information across multiple frames in previous work is that the consecutive hazy frames must be aligned. This can be

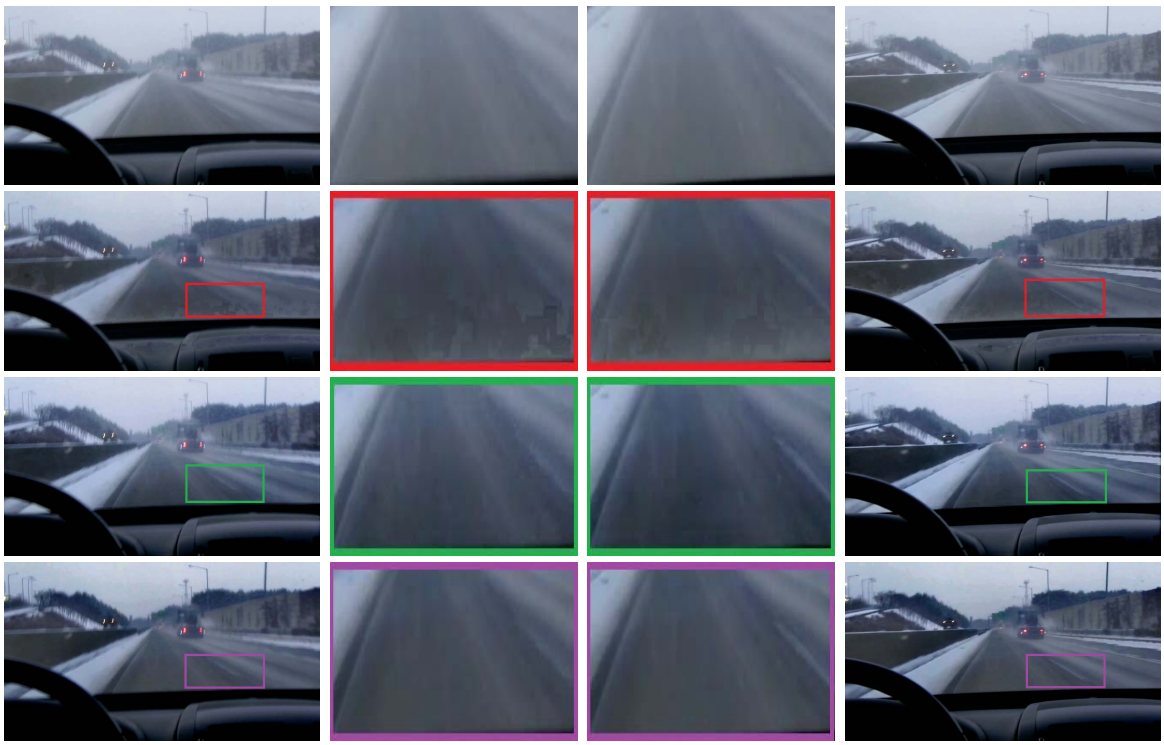


Fig. 1. Video dehazing comparison according to temporal consistency. The first row displays two consecutive hazy input frames and the zoom-in regions. The second and third rows show the dehazed results generated by the method of Zhu *et al*. [23] and DCPDN [24], respectively. The zoom-in regions in these two rows show that the dehazed patterns by the single image dehazing methods [24], [25] are of different appearances between the consecutive frames, which creates flicker artifacts. The last row shows the dehazed results by our VDHNet. As shown, the proposed VDHNet maintains the same appearance of the patterns. (a) Frame t. (d) Frame $t + 1$.

done by using optical flow [21], [22]. However, warping-based alignment is not robust around dis-occlusions and areas with low texture, and often yields warping artifacts.

Recently, great progress has been achieved by applying deep convolutional neural networks (CNNs) to image processing tasks. This kind of tasks includes image denosing [26], super-resolution [27], style transferring [28], and so on. A natural way to extend image processing techniques to videos is to perform a certain image transformation frame by frame. However, this scheme inevitably brings temporal inconsistencies and thus causes severe flicker artifacts. The second and third rows in Figure 1 show the dehazed results by directly applying the learning based image dehazing methods of [23] and [24] to videos, respectively. It can be observed that the zoom-in content marked by red and green rectangles have different appearances between two consecutive frames, therefore creating flicker artifacts. The reason is that slight variations between adjacent video frames may be amplified by the frame-based network and thus result in obviously different dehazed frames. In the existing literature, one solution to retain temporal coherence after video transformation is to explicitly consider temporal consistency during the frame generation or optimization process [29]. While effective, they are case-specific methods and thus cannot be easily generalized to other problems. Among them, the method of Zhang *et al*. [29] is specifically designed for video dehazing. However, it relies on time-consuming optimization about optical flow estimation.

Considering the efficacy of deep networks on image recovering tasks, a natural thinking will be whether CNNs can be adapted to video dehazing tasks by including temporal consistency [30]–[32]. Inspired by the recent work [33], [34] which show that stacked consecutive frames could model temporal information, we consider recovering clear videos by using stacked consecutive frames as input in our proposed network. In this paper, we testify this idea on the problem of video dehazing. We demonstrate that a feed-forward network cannot only capture content information in the spatial domain, but also encourage consistency in the temporal domain. We also present dehazed results with and without alignment before feeding the stacked frames into the network. We show that using the encoder-decoder network with dilated convolution and skip connections can achieve high-quality results without any alignment at all, which makes our approach highly efficient and robust to scene types.

In addition, we introduce a novel semantic segmentation branch which uses semantic information to provide additional guidance for inferring transmission maps. Semantic clues have seen success in other low-level applications, *e.g.* image deblurring [35], image super-resolution [36] and single image dehazing [37]. Here we propose a new convolutional neural network that learns the correlation between semantic segmentations and transmission maps from training samples. If the semantic segmentation of the scene is known, transmission map within the same object should be smooth but the transmission across

the boundary needs not be smooth, such constraints facilitate accurate transmission estimation. An example of dehazed results of our method is shown in the last row of Figure 1, from which we can see that the dehazed results have vivid color information and no more flicker artifacts.

Our work makes three-fold technical contributions:

- A novel haze removal method for videos is proposed, which is solely based on a feed-forward convolutional neural network that exploits global semantic information for video dehazing.
- We demonstrate that a feed-forward convolutional neural network only taking a short stack of neighboring video frames can not only dehaze each video frame well, but also maintain the temporal consistency without any alignment.
- To train the deep network, we create a hazy video dataset using the image sequence and the corresponding depth map from the NYU depth dataset [38]. We compare qualitatively to real videos previously used for video dehazing, and quantitatively with a synthesized dataset with ground truth.

In this paper, we extend our preliminary work [39] in four aspects. First, we investigate the effectiveness of the proposed network with and without using alignment before feeding the consecutive frames into the network (Section III). Second, we exploit the scene semantic information as global priors to better estimate the transmissions for video dehazing (Section III). Third, we add skip connections and dilated convolution in the network and show their better performance for dehazing (Section III). Lastly, we present more technical details, performance evaluation, and analysis in Section V.

II. Related Work

There exist three main approaches for dehazing: clear image priors based methods, approaches that rely on multi-image aggregation or fusion, and CNNs based methods.

A. Dehaze Using Priors

Most existing single-image dehazing approaches [6], [23] jointly estimate a transmission map and the underlying haze-free image via designing clear image priors [40], [41]. For example, Tan *et al*. [42], [43] propose dehazing methods based on two observations. One is that clear images have more contrast than hazy images; the other one is that the variations of atmospheric light, which mainly depends on the distance of objects to the viewer, tend to be smooth. He *et al*. [6] propose a single image dehazing method based on the statistical observation of the dark channel, which allows a rough estimation of the transmission map. Then, they use the expensive matting strategy to refine the final transmission map. Zhu *et al*. [23] find that the difference between brightness and saturation in a clear image should be very small. Therefore, they propose a new color attenuation prior based on this observation for haze removal from a single input hazy image. Recently, Berman *et al*. [8] introduce a non-local method for single image dehazing. This approach is based on the assumption that an image can be faithfully represented with just a few hundreds of distinct colors.

All of the above approaches strongly rely on the accuracy of the assumed image priors, thus may perform poorly when the hand-crafted priors are insufficient to describe real data. As a result, these approaches tend to be more fragile than aggregation-based methods [17], and often introduce undesirable artifacts such as amplified noises.

B. Multi-Image Aggregation/Video Dehazing

Multi-image aggregation methods directly combine multiple images in either spatial or other domains (e.g., chromatic, luminance and saliency) without solving any inverse problem by retaining only the most useful features. Most existing works merge multiple low-quality images into the final result [1], [17], [20]. Kim *et al*. [20] assume that a scene point yields highly correlated transmission values between adjacent image frames, then add the temporal coherence cost to the contrast cost and the truncation loss cost to define the overall cost function. However, the pixel-level processing increases the computational complexity, therefore, may not be suitable for handling videos. Choi *et al*. [44] first make use of measurable deviations from statistical regularities observed in natural foggy and fog-free images to predict fog density, then develop a referenceless perceptual image defogging algorithm based on estimated fog density. Zhang *et al*. [29] first dehaze videos frame by frame, and then use optical flow to improve the temporal coherence based on Markov Random Field (MRF).

However, all the above approaches have explicit formulations on how to fuse multiple images. In this work, we instead adopt a data-driven approach to learn how multiple images should be aggregated to generate transmission maps.

C. Data-Driven Approaches

Recently, CNNs have been applied to achieve leading results on a wide variety of reconstruction problems. These methods tend to work best when large training datasets are easily constructed [45], [46]. Such as image denoising [26], image deraining [47], [48], and super-resolution [27]. However, these approaches address a different problem, with its own set of challenges. In this work we focus on video dehazing, where neighboring hazy frames can provide abundant information for transmission map estimation.

CNNs have also been used for single image dehazing based on synthetic training data [49]. Unlike traditional methods that use hand-crafted priors to estimate the transmission map, Ren *et al*. [50] first use a coarse-scale CNN to extract a holistic transmission map, then propose a fine-scale CNN to refine the output from coarse-scale CNN. Cai *et al*. [51] present a DehazeNet and a Bilateral Rectified Linear Unit (BReLU) for transmission estimation. Recent AOD-Net [52] bypasses the transmission estimation step by introducing a new variable $\mathbf{K}(x)$ which integrates both transmission $t(x)$ and atmospheric light $\mathbf{A}$. However, these algorithms focus on static image dehazing and inevitably yield flickering artifacts due to the lack of temporal coherence when applied to video dehazing. Li *et al*. [3] extend the AOD-Net and propose a EVD-Net for video dehazing by considering the temporal coherence between neighboring video frames. This method effectively

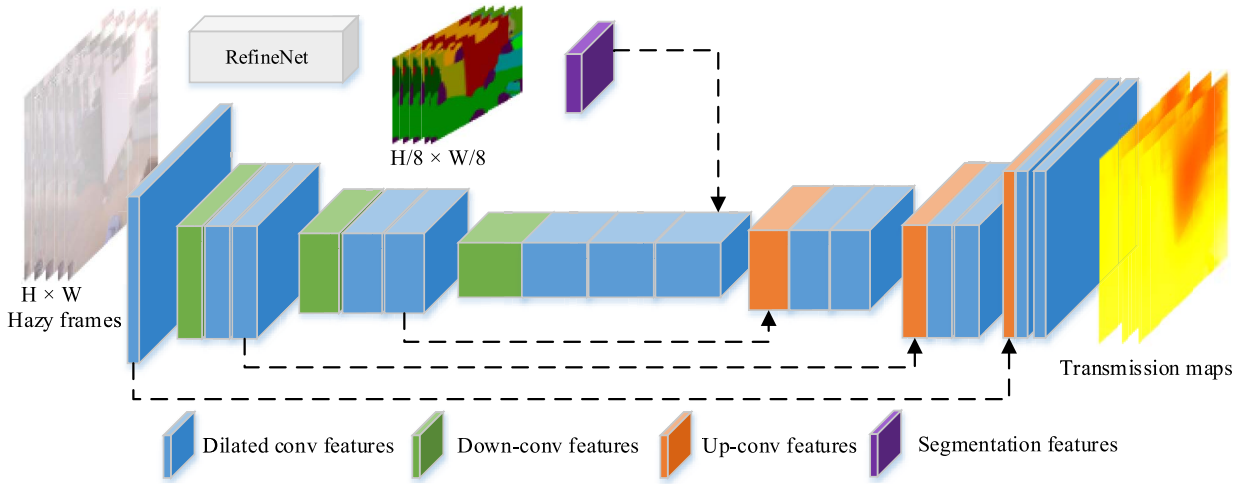


Fig. 2. Our proposed VDHNet. The proposed encoder-decoder style architecture that takes in five consecutive hazy frames stack and produces central three transmission maps. We use the features of semantic information to better estimate the transmission maps.

maintains the coherence of consecutive frames, but tends to remain some haze in the dehazed results.

D. Semantic Segmentation

Semantic segmentation aims to cluster image pixels of the same object class with assigned labels. Numerous recent methods use semantic segmentation to resolve image deblurring [35], super-resolution [36] and single-image dehazing [37].

In our experiments, we show that multi-frame transmission maps estimation and the temporal consistency can be simultaneously learned by leveraging multiple video frames and our proposed deep network, which avoids computing optical flows in the inference stage and thus enables real-time dehazing for videos. In addition, we show that semantic segmentation provide informative priors for estimating transmission maps.

III. Our Method

In this section, we train an end-to-end system for video dehazing, where the input is a stack of five neighboring frames and the output are the estimated transmission maps of central three frames in the stack. The estimated transmission maps are subsequently used to recover a haze-free video via atmospheric scattering model. We use the stacked frames with and without alignment before feeding the frames into the network. In addition, we exploit the semantic information from a semantic segmentation network [53] as the global semantic priors within the deep network. In the following, we first present our neural network architecture, and then describe a number of experiments for evaluating its effectiveness and comparing with existing dehazing methods [3], [6], [24], [51], [52], [54], [55]. The key advantage of our method is the allowance of lessening the requirements for accurate alignment, a fragile component of prior work, but implicitly enforces the temporal coherence of estimated transmission maps of neighboring frames.

A. Network Architecture

We use an encoder-decoder style network, which has been shown to produce good results for a number of generative tasks.

Fusion Strategy: To keep the temporal coherence of neighboring transmission maps, we use two different strategies to fuse consecutive frames. The input of the network is five consecutive frames (f_{t-2}, f_{t-1}, f_t, f_{t+1} and f_{t+2}) with different fusion strategies. Note that any number of past and future frames can be accommodated in the input layer. In order to use more than one forward- and backward-frame, the architectures in Figure 2 can be directly extended by fusing more frames according to channel dimension.

1) Fusion Without Alignment: First, we perform an early fusion of neighboring hazy frames by concatenating five consecutive images in the input layer as shown in Figure 2. Then, all the five concatenated frames are fed into the first dilated convolutional layer without any alignment at all, relying on the network to extract spatial and temporal information through a series of dilated convolutional layers. The output of the network is the central three transmission maps of the neighboring five hazy images. This makes the network significantly faster since alignment usually dominates running time in multi-frame aggregation methods. We refer to this network as VideoDehazeNet, or VDHNet.

2) Fusion With Alignment: Then, we conduct a fusion strategy by using alignment before feeding the consecutive frames into the network. We use optical flow [56] to align stacked frames, which is time-consuming and tend to introduce additional warping artifacts [21], but allows pixels to be aggregated more easily by removing the spatial variance of corresponding features. Let f_t be the target frame, f_{t-2}, f_{t-1}, f_{t+1} and f_{t+2} are the frames need to be warped. We first estimate optical flows between each frame and the target, and then proceed to interpolate the set of consistent pixels [57]. The wrapped frames are concatenated and then feed to the encoder-decoder network. Different from the fusion strategy without alignment, the output of the network is the central transmission map

TABLE I

SPECIFICATIONS OF THE PROPOSED VDHNET MODEL. AFTER EACH CONVOLUTIONAL AND DECONVLUTIONAL LAYER, EXCEPT THE LAST ONE, THERE IS A RECTIFIED LINEAR UNIT. WE PAD ALL CONVOLUTIONAL LAYERS WITH ZEROS SUCH THAT THE OUTPUT SIZE IS THE SAME AS THE INPUT SIZE WHEN USING A STRIDE OF 1. ALL OUTPUT SIZES REFER THE ORIGINAL IMAGE WIDTH W AND HEIGHT H, AS THE MODEL CAN PROCESS IMAGES OF ANY RESOLUTION

Layer	Kernel size	Dilated factor	Stride	Output size	Skip connection
Input	-	-	-	$15 \times H \times W$	-
RefineNet	-	-	-	$5 \times H \times W$	-
Conv0+Pooling	-	-	-	$64 \times H/8 \times W/8$	to Dilated conv 4-3
Dilated conv 1-1	5×5	-	1×1	$64 \times H \times W$	to Up conv 3
Dilated conv 2-1	3×3	2	2×2	$128 \times H/2 \times W/2$	-
Dilated conv 2-2	3×3	2	1×1	$128 \times H/2 \times W/2$	-
Dilated conv 2-3	3×3	2	1×1	$128 \times H/2 \times W/2$	to Up conv 2
Dilated conv 3-1	3×3	2	2×2	$256 \times H/4 \times W/4$	-
Dilated conv 3-2	3×3	2	1×1	$256 \times H/4 \times W/4$	-
Dilated conv 3-3	3×3	2	1×1	$256 \times H/4 \times W/4$	to Up conv 1
Dilated conv 4-1	3×3	2	2×2	$512 \times H/8 \times W/8$	-
Dilated conv 4-2	3×3	2	1×1	$512 \times H/8 \times W/8$	-
Dilated conv 4-3	3×3	2	1×1	$512 \times H/8 \times W/8$	from RefineNet
Up conv 1	4×4	-	$1/2 \times 1/2$	$256 \times H/4 \times W/4$	from Dilated conv 4-3
Dilated conv 5-1	3×3	2	1×1	$256 \times H/4 \times W/4$	-
Dilated conv 5-2	3×3	2	1×1	$256 \times H/4 \times W/4$	-
Up conv 2	4×4	-	$1/2 \times 1/2$	$128 \times H/4 \times W/4$	from Dilated conv 3-3
Dilated conv 6-1	3×3	2	1×1	$128 \times H/4 \times W/4$	-
Dilated conv 6-2	3×3	2	1×1	$128 \times H/4 \times W/4$	-
Up conv 3	4×4	-	$1/2 \times 1/2$	$64 \times H/2 \times W/2$	from Dilated conv 2-3
Dilated conv 7-1	3×3	2	1×1	$64 \times H/2 \times W/2$	-
Dilated conv 7-2	3×3	2	1×1	$64 \times H/2 \times W/2$	-
Output	-	-	-	$3 \times H \times W$	

(*i.e.,* transmission of frame f_t) of the neighboring hazy images in this section. We refer to the network as VDHNet+alignment.

Both the training losses of the two fusion strategies are the MSE to the synthetic ground truth transmission maps as

$$L(t_i(x), t_i^*(x)) = \frac{1}{pq} \sum_{v=1}^{p} \sum_{f=1}^{q} ||t_{v,f}(x) - t_{v,f}^*(x)||^2, \quad (2)$$

where f is the frame index and v denotes the video index, p and q is the number of training videos and frames in each video, respectively. Both networks consist of three types of layers: down-convolutional layers, that compress the spatial resolution of the features while increasing the spatial support of subsequent layers; the up-convolutional layers, *i.e.*, deconvolutional layers, that increase the spatial resolution; and convolutional layers. We use convolutions after both down-convolutional and deconvolution layers to further sharpen the activation results. Note that we use dilated convolution in all the convolutional layers.

a) Dilated convolution: For haze removal task, contextual information from an input image is demonstrated to be useful for automatically identifying and removing the haze. The dilated convolution [58] weights pixels with a step size of the dilated factor, and thus increases its receptive field without losing resolution. Thus, we propose a contextualized dilated network to aggregate context information for learning the haze relevant features since it provides an increasingly larger receptive field for the following layers. We use the same dilated factors in each layer in the proposed network. Specifically, the dilated factor is 2 in all the layers. Please refer to Table I for detailed configurations of the network.

b) Skip connection: We also use the skip links to guide the estimation of transmission maps. We note that the encoder process removes the details but preserves the main structures from the input images while decoder process concatenates the features from encoder process and the features from the shallow layers (whose features contain edges and finer details information) to generate more useful features for transmission estimation. The skip link is added on every scale to utilize the feature maps in the encoder and help maintain the details from the encoders.

c) Global semantic prior: We propose to utilize the semantic segmentation information as a global prior for video dehazing. Ideally, the transmission map should be smooth in the regions of the same object and discontinuous across the boundaries of different objects. Therefore, we expect the estimated transmission map to be smooth inside the same object, and discontinuous only along depth edges. As such, we propose a new semantic information branch by using the RefineNet [53]. Since the RefineNet learns rich semantic representations for input images, it is important to resolve the ambiguity in edge and object boundary in transmission maps. Thus, given the consecutive frames, we first use the semantic segmentation network [53] to extract the semantic labels. We then extract semantic features from the probability maps of the semantic labels using an additional convolutional layer as shown in Figure 2. Finally, we concatenate the semantic features with the features of hazy frames as the input to the first decoder block.

Here we demonstrate the robustness of the semantic segmentation network [53]. Given a haze-free image in Figure 3(a), we use different medium extinction coefficient β to synthesize images with different haze concentration. As shown in the second row of Figure 3(b)-(e), all the segmentations estimated by the RefineNet [53] are similar

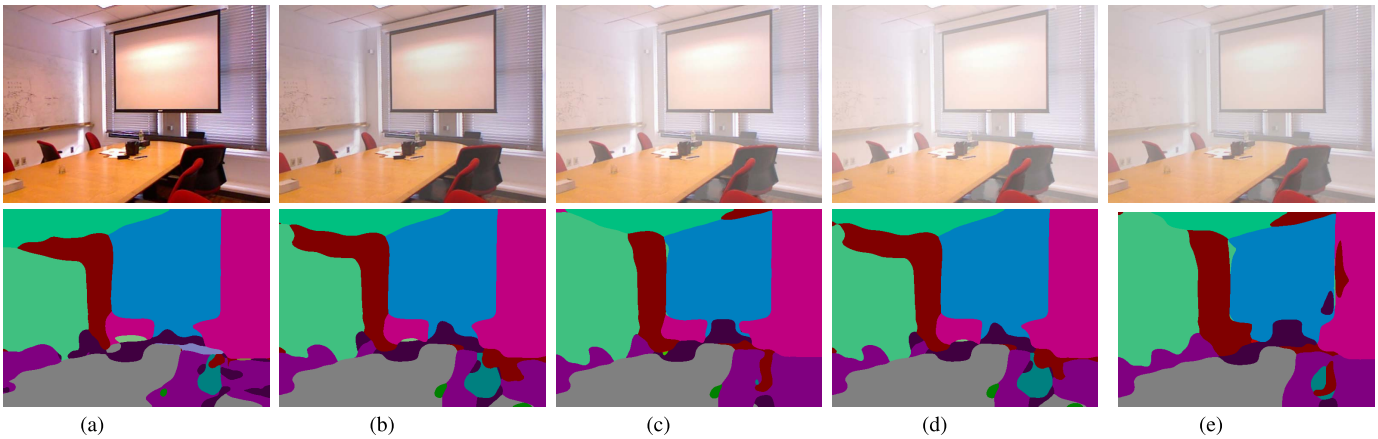


Fig. 3. Semantic segmentation by the Refinenet [53]. Top: Clear image and synthetic hazy images with different medium extinction coefficient β. Bottom: Semantic segmentations estimated by the Refinenet [53]. All of these segmentations estimated by the Refinenet [53] are similar for different haze concentrations. (a) Haze-free. (b) $\beta = 0.6$. (c) $\beta = 0.8$. (d) $\beta = 1.0$. (e) $\beta = 1.2$.

for different haze concentration images. We set the size of semantic segmentations as 1/8 of the input hazy frames since low-resolution images could mitigate the mismatch of edges between segmentations and hazy frames. Therefore, there exists a total of 576-channel (512-channel features of RGB images and 64-channel features of semantic probabilities) features before the first up-convolutional layer. These 576-channel features are then up-sampled by $2\times$ through a deconvolutional layer. Figure 2 shows an overview of our video dehazing network. The semantic labels encode the essential appearance information and serve as a strong global prior for reconstructing the transmission maps. Note that we use the models of "RefineNet-NYUD" and "RefineNet-Cityscapes" to segment the indoor and outdoor images, respectively.

With the designed VDHNet, we estimate transmission maps directly from input hazy fames. After estimating the transmission maps for the inputted video frames, we use the method in Section IV to compute atmospheric light and recover final dehazed frames based on the atmospheric scattering model (1).

B. Implementation Details

During training we use a batch size of 5, and patch size of $15 \times 256 \times 256$, where 15 is the total number of channels stacked from the crops of 5 consecutive video frames. We observed that a patch size of 256 was sufficient to provide enough overlapping content in the stack even if the frames are not aligned, which has also been reported in [21]. We use ADAM for optimization, and initialize the learning rate as 0.00001. For all the results reported in the paper we train the network for 100,000 iterations, which takes about 80 hours on an NVidia K80 GPU. Default values of β_1 and β_2 are used, which are 0.9 and 0.999 respectively, and we set weight decay to 0.00001. Since our approach dehazes frames in a single forward pass, it is computationally very efficient. Using an NVidia K80 GPU, we can process three 640×480 frames within 0.2s. Previous approaches took on average 26s [6] and 3s [54] per frame on CPUs. The recent video dehazing method [55] takes more than 100s for each frame.

C. Training Dataset

Generating realistic training data is a major challenge for video dehazing task in which ground truth data cannot be easily collected [59]. To train our deep network, we need to generate a dataset with synthesized hazy videos and their corresponding transmission maps and haze-free videos. Although we can use the state-of-the-art method [60] to add synthetic haze to clear outdoor scenes, we found the synthesized indoor hazy images also generalize well to outdoor scenes [50]. Therefore, we use the NYU Depth dataset [38] to synthesize training data in this paper. The NYU Depth dataset [38] contains 587 video clips and corresponding depth values. We randomly sample 100 clean video clips and the corresponding depth maps to construct the training set. Given a clear frame $\boldsymbol{J}(x)$ and the ground truth depth $d(x)$, we synthesize a hazy image using the physical model (1). We generate the random atmospheric light $\boldsymbol{A} = [k, k, k]$, where $k \in [0.8\ 1.0]$. We synthesize transmission map for each frame based on the ground truth depth map $d(x)$,

$$t(x) = e^{-\beta d(x)}. \tag{3}$$

Mathematically, β and depth $d(x)$ in (3) have equal effects on the generated transmission map. We sample four random $\beta \in [0.5\ 1.5]$ in (3) for every video clip. Therefore, we have 400 hazy videos and the corresponding haze-free videos (100 videos $\times$ 4 medium extinction coefficients β) in the training set. Examples of the dataset are shown in Figure 4.

IV. Video Dehazing With VDHNet

In addition to transmission maps $t(x)$ that is learned from the proposed VDHNet, we need to estimate the atmospheric light $\boldsymbol{A}$ for each frame in order to recover the clear video according to (1). We compute $\boldsymbol{A}$ from the estimated transmission map as like in [50] and [54] but with the consideration of smoothness. Given a single frame, we estimate the atmosphere light $\boldsymbol{A}$ by giving a threshold t_{thre},

$$\boldsymbol{I}(x) = \boldsymbol{A}, \quad t(x) < t_{thre}. \tag{4}$$

According to (4), we select 0.1% darkest pixels in a transmission map $t(x)$. These pixels have the most haze concentration.

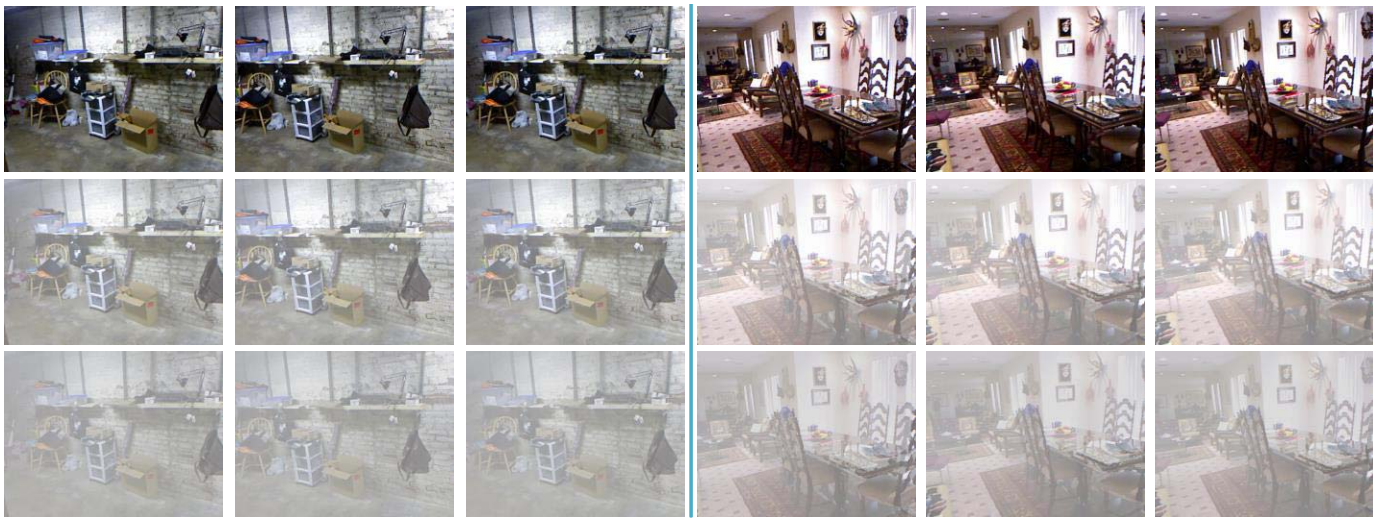


Fig. 4. Synthetic hazy videos for training. The top row is the haze-free frames of two videos from the NYU Depth dataset [38], and the bottom two rows are our synthetic hazy frames with different haze thickness.

TABLE II
AVERAGE MSE OF THE ESTIMATED TRANSMISSION MAPS ON THE 5 SYNTHETIC VIDEOS AND THE AVERAGE RUNNING TIME OF EACH FRAME

Method	VDHNet	VDHNet+alignment
MSE	0.0417	0.0421
Average time (s)	0.12	35.94

Among these pixels, the one with the highest intensity in the corresponding hazy image $\boldsymbol{I}$ is selected as the atmospheric light. But in addition, we use the average value of the estimated three consecutive lights as the final atmospheric light in our experiment to maintain the smoothness.

After the atmospheric light $\boldsymbol{A}$ and the scene transmission map $t(x)$ of each frame are estimated by the proposed VDHNet, we recover the haze-free frames by reversing the model (1).

V. EXPERIMENTAL RESULTS

In this section, we conduct a series of experiments to evaluate the effectiveness of the learned model.

A. Alignment

We first compare the performances of two different fusion strategies with and without alignment in our VDHNet. We use 5 synthetic video clips to test the accuracy of transmission estimation. As shown in Table II, both two fusion strategies have similar transmission estimation accuracy. In this paper, we choose the VDHNet without alignment as the default configuration since these two networks only have 0.0004 distinctions according to MSE of the estimated transmission maps. In addition, the network without alignment makes the VDHNet significantly faster since alignment usually dominates running time as shown in Table II. The average running time of the network with alignment is 35.94s. In contrast, the average running time of the fusion strategy without alignment is only 0.12s. Therefore, all the results reported in this work are based on the VDHNet without using alignment.

B. Effectiveness of the Skip Links

The skip connections act as a guidance operation which concatenates the different scale features in the VDHNet. Without the skip links, the network are not able to learn the features with fine details for image dehazing, which accordingly affects the details estimated in the dehazed results (see Figure 5(b)).

We note that the features for early layers usually contains finer details. Thus, we adopt skip links to use these features from shallow layers for the details estimation of transmission maps. Compare with the network without skip links in Figure 5(b), our network with skip links generates much better results with finer details as shown in Figure 5(d).

C. Effectiveness of the Dilated Convolutions

Contextual information from an input degraded image has bee demonstrated to be useful for automatically image recovering [61]. Thus, we exploit a dilated network to aggregate context information for learning the haze-relevant features, which provides an increasingly larger receptive field for the following layers.

In Figure 5(c), we show the result generated by the network without using dilated convolutions. Note that the color of the building in the red rectangle tends to be darker than it should be. In contrast, the network with dilated convolution yields better visual result as shown in Figure 5(e).

To further show the effectiveness of the proposed dilated convolution layers and skip connections, we show the quantitative results with different configurations in Table III. As can be seen, our network with skip connection and dilation convolutions could generate better results according to PSNR on the synthetic videos.

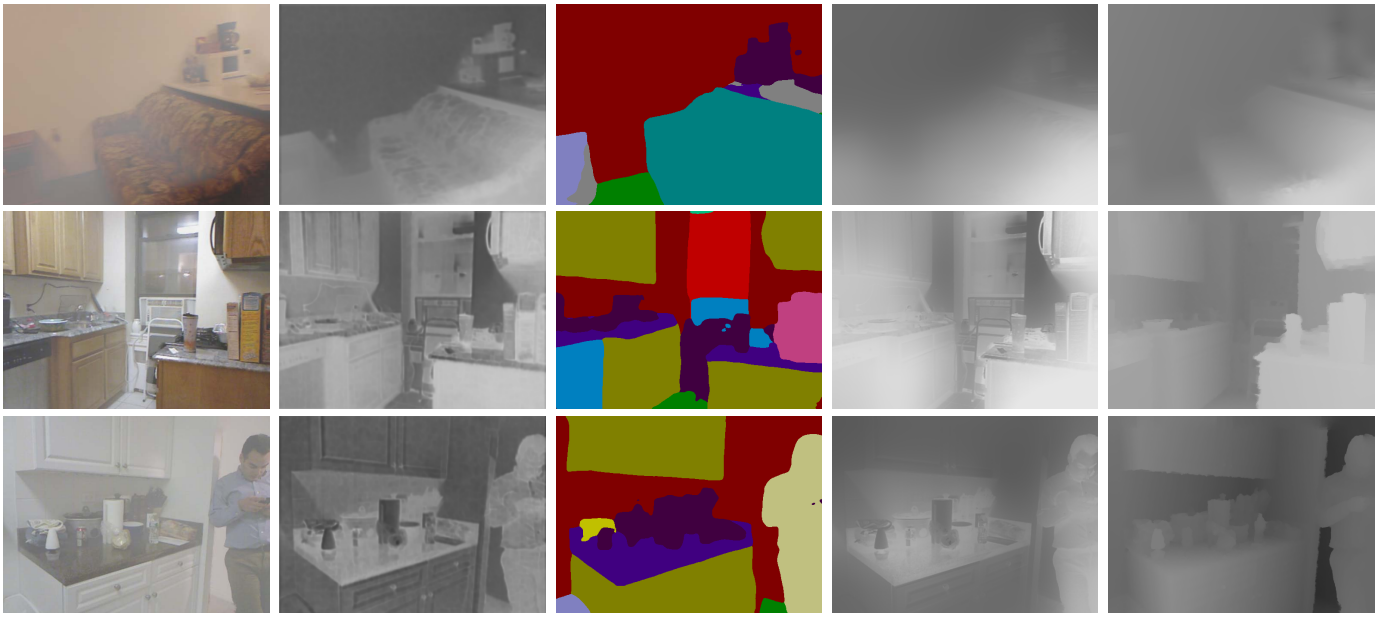
(a) (b) (c) (d)

Fig. 5. Visual comparison using different configurations. (a) Hazy input. (b) and (c) show the results without skip connections and dilation convolution layers, respectively. (d) Our result.

TABLE III

AVERAGE PSNR OF THE RECOVERED IMAGES ON THE 5 SYNTHETIC HAZY VIDEOS

Method	w/o skip and dilation [41]	w/o skip	w/o dilation	w/o segmentation	Ours
PSNR	23.45	23.56	23.87	23.61	**23.95**



(a) (b) (c) (d) (e)

Fig. 6. Effectiveness of semantic segmentation for transmission map estimation. (a) Hazy frames. (b) Estimated transmission maps by the VDHNet without segmentation. (d) Estimated transmission maps with the semantic information in (c). (e) Ground-truth transmission maps. As shown, our estimated transmission maps are smooth in the same object regions and only discontinuous across the boundaries of different objects.

D. Effectiveness of Semantic Segmentation

Semantic segmentation improves video dehazing in multiple ways as it is used to help estimate transmission maps from which the haze-free frames are estimated. First, it provides region information about object boundaries. Second, as different objects have different depth maps, semantic segmentations are used to constrain transmission maps estimation of each region. As shown in the first row of Figure 6(b), the estimated transmission map has some extraneous edges in the "sofa" region when the semantic segmentation is not used. In contrast, the semantic information in (c) helps generate accurate transmission map as shown in Figure 6(d). The estimated transmission is smooth inside the same object, and discontinuous only along depth edges. The quantitative results in Table III also demonstrate the effectiveness of the proposed semantic segmentation branch.

E. Temporal Smoothness

In this section, we use the temporal error $E_{temporal}$ [28] to evaluate the temporal smoothness of the dehazed videos by different models and compare with the state-of-the-art methods of [3], [51], and [52]. Table V lists the temporal errors of five different models on the testing videos synthesized by the NYU-Depth dataset. It shows that the proposed VDHNet achieves the smallest temporal error, the video dehazing model EVD-Net follows, and the DCPDN [24] turns out to be least temporally coherent.

TABLE IV

AVERAGE PSNR AND SSIM OF DEHAZED RESULTS ON THE 20 SYNTHETIC VIDEOS

	He et al. [19]	Chen et al. [55]	MSCNN [52]	DehazeNet [53]	EVD-Net [6]	Our VDHNet
PSNR	19.79	20.28	21.94	22.01	22.50	**22.93**
SSIM	0.67	0.66	0.81	0.80	0.82	**0.85**

TABLE V

TEMPORAL ERRORS OF THREE DIFFERENT METHODS ON 5 TESTING VIDEOS

DehazeNet [53]	AOD-Net [54]	DCPDN [27]	EVD-Net [6]	VDHNet
0.0934	0.0520	0.0971	0.0438	0.0414

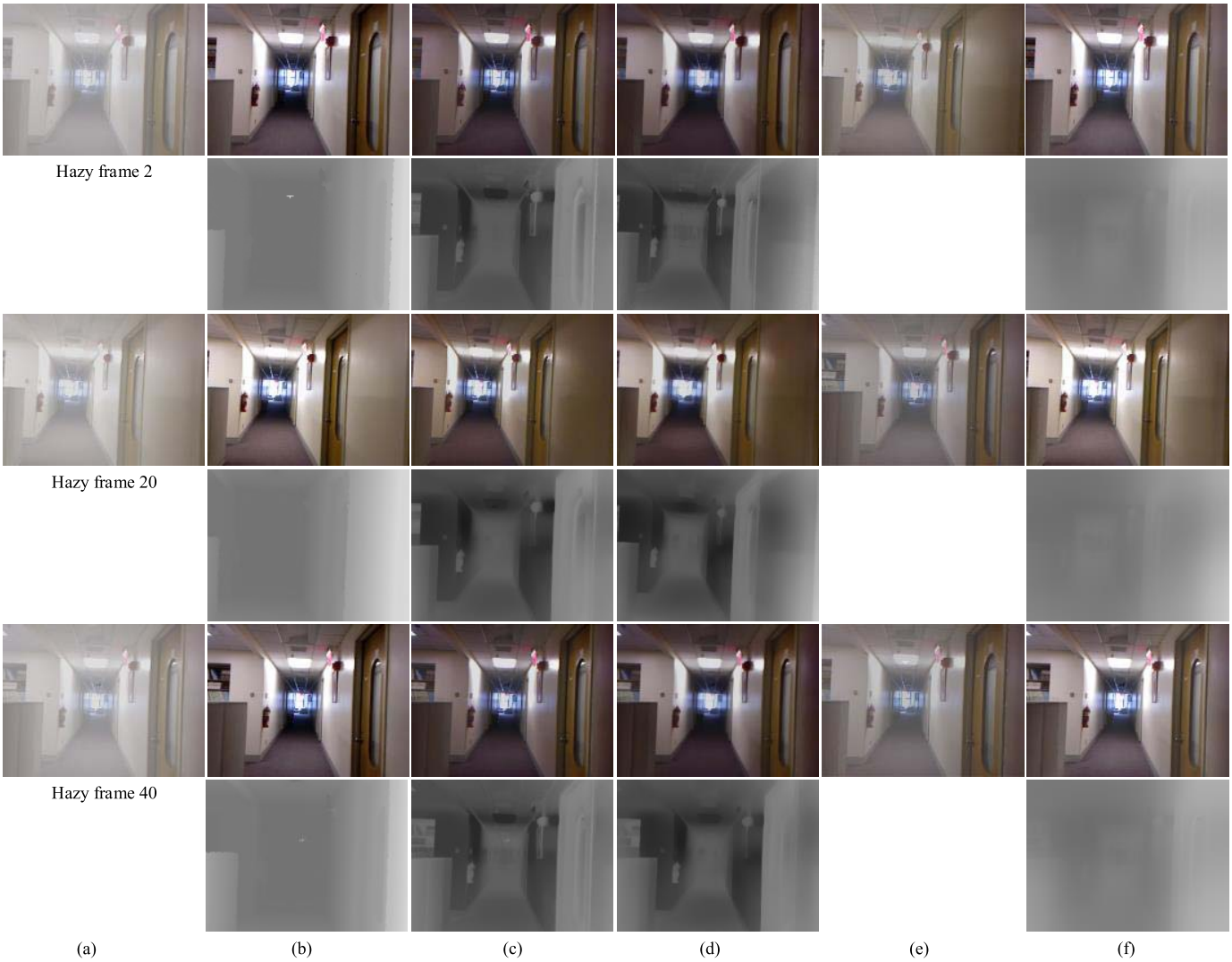


Fig. 7. Dehazed results (odd rows) and estimated transmission maps (even rows) on our synthetic videos. (a) Input hazy frames. (b) Ground-truth, (c) He *et al*. [6], (d) Chen *et al*. [55], (e) AOD-Net [52] and (g) Our VDHNet. Both single image dehazing method [6] and the recent video dehazing method [55] tend to over-estimate haze concentration and result in dark results. In contrast, our approach performs better than the state-of-the-art methods.

F. Quantitative Evaluation

In this section, we compare our proposed VDHNet with the state-of-the-art single image dehazing methods of He *et al*. [6], Ren *et al*. [50], Cai *et al*. [51] and AOD-Net [52], and compare with the recent video dehazing methods of Chen *et al*. [55] and EVD-Net [3]. For quantitative performance evaluation, we construct a new testing hazy video dataset. We select 20 video clips and their depth maps from the NYU Depth dataset [38] (different from those that used for training) to synthesize 20 transmission map sequence and corresponding hazy videos. The results of single image dehazing methods of [6] and [50]–[52] and the video dehazing approach [55] are from the authors' implementations and we used the hand-tuned parameters to produce the best possible results. For the

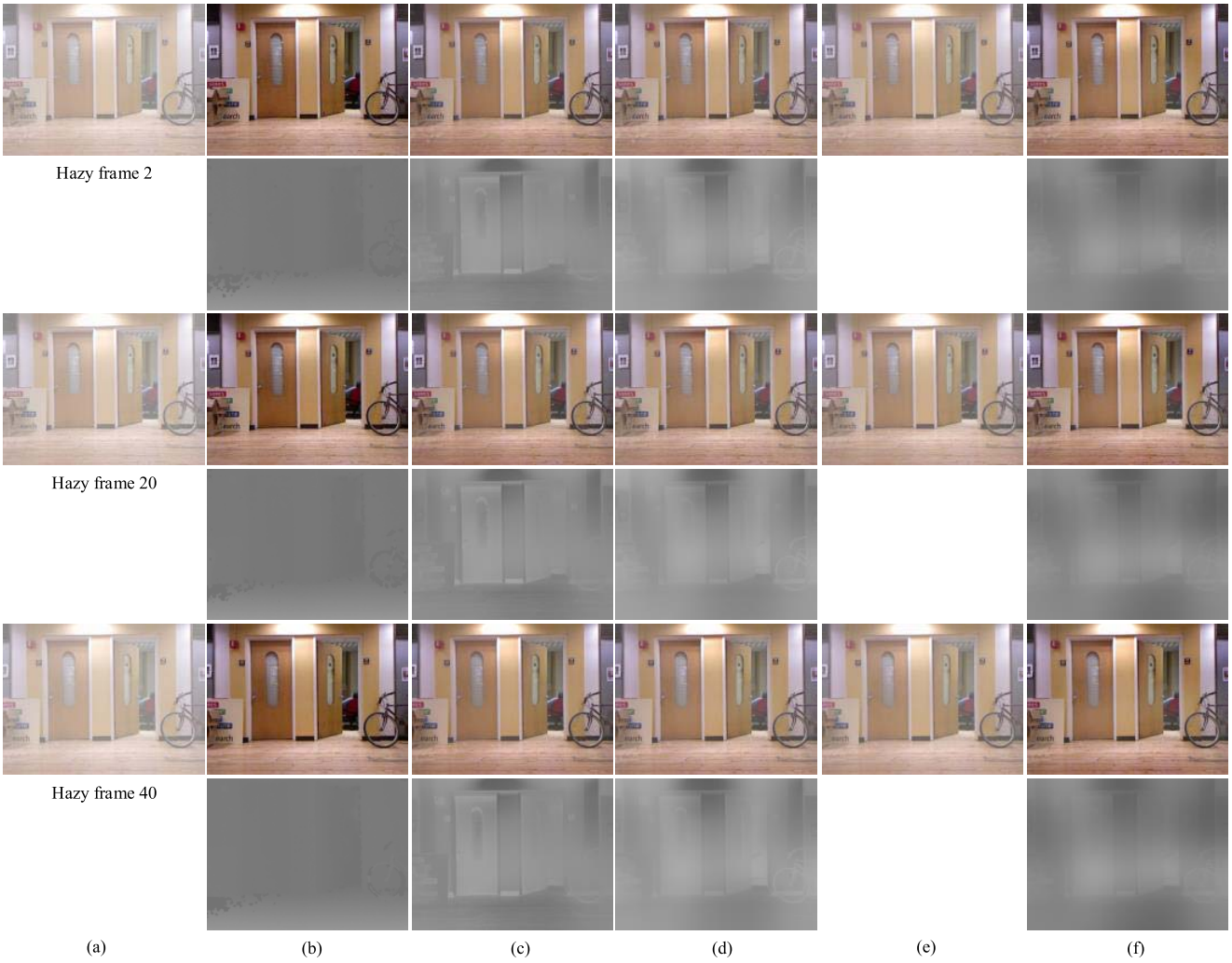


Fig. 8. Dehazed results (odd rows) and estimated transmission maps (even rows) on our synthetic videos. (a) Input hazy frames. (b) Ground-truth, (c) Cai *et al*. [51], (d) Ren *et al*. [50], (e) EVD-Net [3] and (f) Our VDHNet method. The CNNs-based methods [51] and [50] tend to transfer extraneous textures to the estimated transmission maps. In contrast, our approach performs better than the state-of-the-art methods.

video dehazing method [3], we re-train the EVD-Net using the same training set in our work.

Figure 7 shows some dehazed frames by the methods [6], [52], and [55]. The estimated transmission maps by He *et al*. [6] are overestimated in some slight hazy regions. Therefore, the dehazed results tend to be darker than the ground truth images in some regions, e.g., the ceiling in Figure 7(c). We note that the dehazed results in Figure 7(d) by the video dehazing method [55] are similar to those by He *et al*. [6] in Figure 3(c) since the method [55] also use the dark channel prior. The final dehazed frames are darker than the ground truth in Figure 7(b). The dehazed results by the AOD-Net still have some remaining haze as shown in Figure 7(e). Figure 7(f) shows the estimated transmission maps and the final recovered images by the proposed VDHNet. Overall, the dehazed results by the proposed algorithm have higher visual quality and fewer color distortions.

In addition, we also compare our algorithm with the learning-based methods [3], [50], [51] in Figure 8. As shown in Figure 8(c) and (d), the methods [51] and [50] are likely to transfer extraneous textures to the transmission map, thereby introducing unnecessary details to the estimated transmission maps. For example, the estimated maps by [51] and [50] contain much texture details in the door area which have the same depth. Although the EVD-Net bypass the transmission map estimation, the dehazed results in Figure 8(e) still have some remaining haze. In contrast, the transmission maps generated by our VDHNet have the similar values in the same depth region. The qualitative results are also reflected by the quantitative PSNR and SSIM metrics shown in Table IV.

G. Real Videos

Although our proposed VDHNet is trained on synthetic indoor videos, we note that it can be applied to outdoor images as well since we could increase the haze concentration by adjusting the value of the medium extinction coefficient β in (3). Therefore, our synthesized transmission maps cover the range of values in real transmission maps as illustrated in [50].

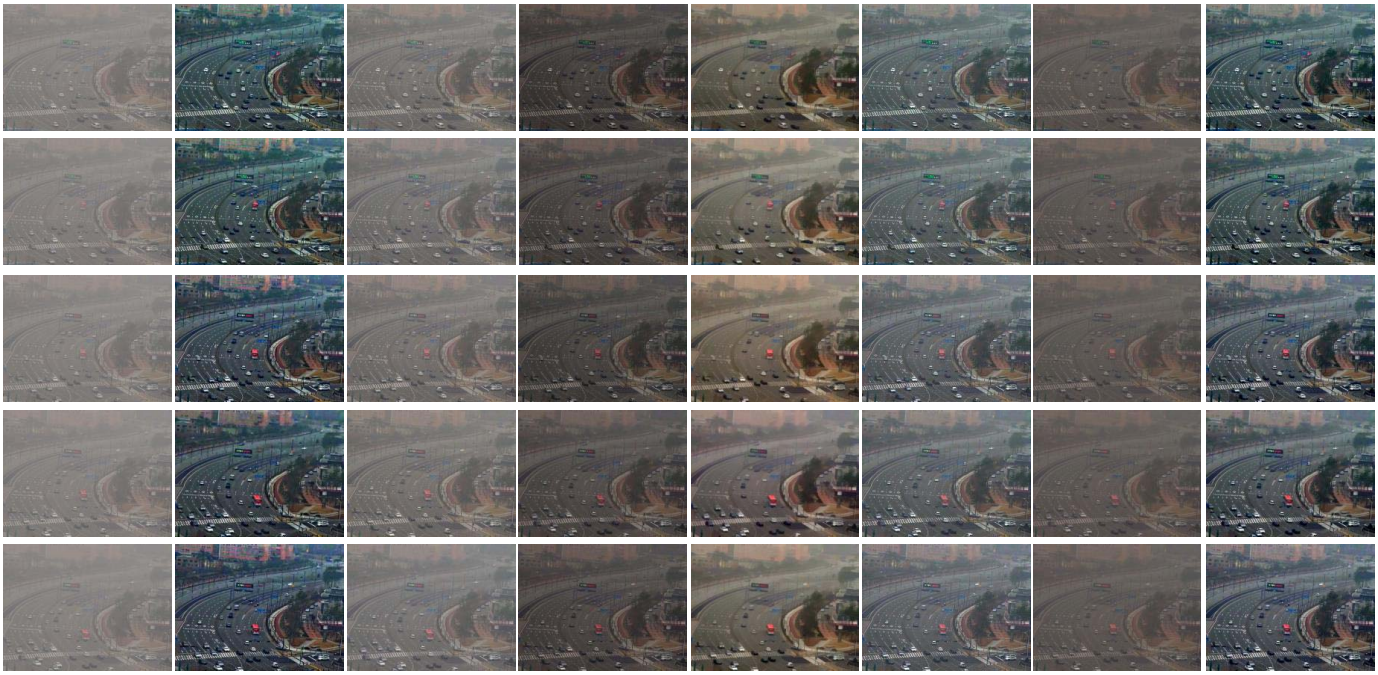


Fig. 9. Dehazed results on the *crossroad* video. (a) Input hazy image. (b) He *et al*. [16], (c) Zhu *et al*. [23], (d) Cai *et al*. [51], (e) Chen *et al*. [55], (f) Ren and Cao [39], (g) EVD-Net [3] and (h) Our VDHNet method. The method of He *et al*. [16] tends to over-estimate the transmission and generate darker results than our algorithm, while the dehazed results by [3], [23], [51], and [55] still have some remaining haze.

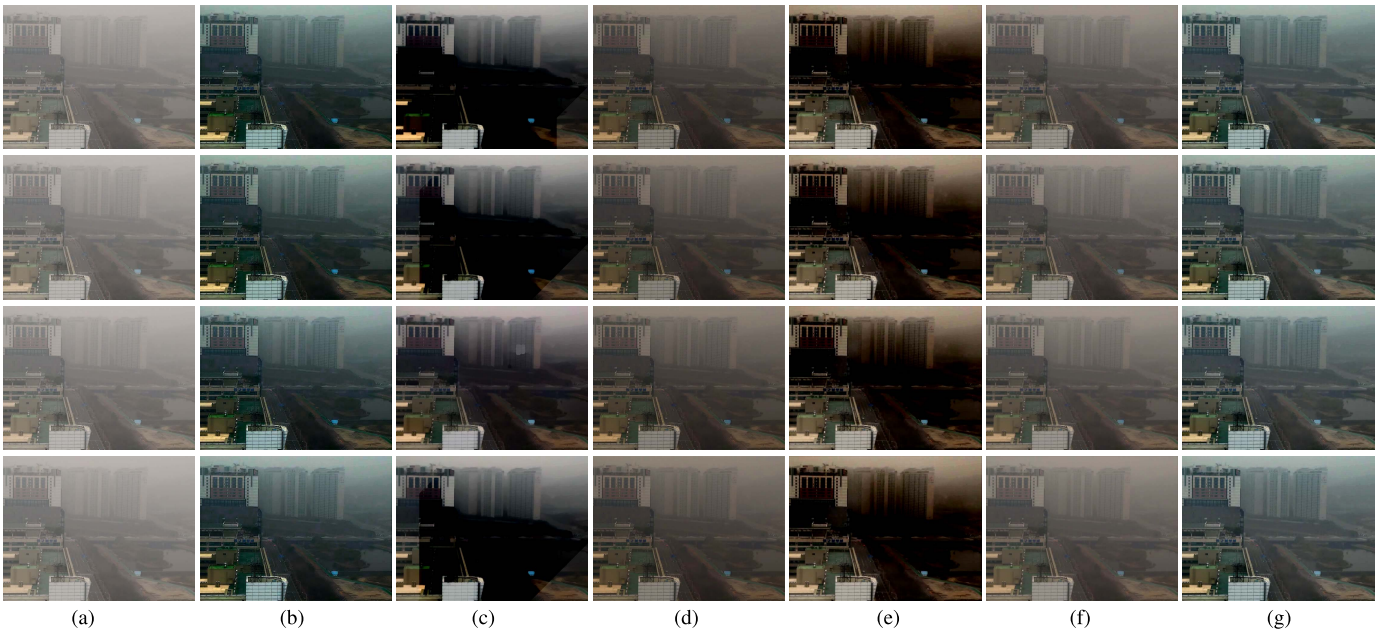


Fig. 10. Dehazed results on the *riverside* video. (a) Input hazy image. (b) Ren *et al*. [50], (c) Chen *et al*. [55], (d) AOD-Net [52], (e) DCPDN [24], (f) EVD-Net [3] and (g) Our VDHNet method. The methods of [24] and [55] tend to generate dark regions in the dehazed results.

We evaluate the proposed algorithm against the state-of-the-art image and video dehazing methods He *et al*. [6], Zhu *et al*. [23], Cai *et al*. [51], Chen *et al*. [55], AOD-Net [52], DCPDN [24], EVD-Net [3] and our previous work of [39] using challenging real videos as shown in Figures 9 and 10. In Figure 9, the dehazed frames by He *et al*. [6] tend to be dark since this method overestimates transmission maps as mentioned in Section V-F. The dehazing methods of Zhu *et al*. [23], Cai *et al*. [51] and Li *et al*. [3] tend to under-estimate the thickness of the haze. Thus, the dehazed videos still have some remaining haze as shown in Figures 9(c), (d) and (f). The method by Chen *et al*. [55] can enhance the image visibility. However, the dehazed results are over-smoothed and many fine details in the recovered images are removed. For example, the *distant buildings* are over-smoothed.

In Figure 10, the methods of Chen *et al*. [55] and Zhang and Patel [24] generate dark results as shown in (c) and

(e). While the methods of MSCNN [50], AOD-Net [52] and EVD-Net [3] tend to remain some haze in the dehazed results as shown in Figure 10 (d) and (f). In contrast, the dehazed results by the proposed VDHNet in Figures 9-10(g) are visually more pleasing in dense haze regions without color distortions or artifacts.

VI. Conclusions and Future Work

In this paper, we proposed a novel deep learning approach for video dehazing in an encoder-decoder style. We show that the temporal coherence of consecutive transmission maps can be automatically learned in a trainable end-to-end system without special design. Our method generates results that are as good as or superior to the state-of-the-art dehazing approaches, without parameter tuning or the requirement of challenging image alignment. We also exploit the semantic information as constraints to better restore the shape of transmission maps of hazy frames.

Although we successfully applied the VHDNet for consecutive transmission maps estimation, we use the indoor training set in this work. In the future, we will train the VDHNet using outdoor images synthesized by the method of [60], and exploit more efficient temporal constraints to further improve the performance of the proposed algorithm. In addition, the semantic branch used in this paper is the pre-trained RefineNet. Therefore, we will design a network for estimating semantic segmentation and transmission maps in a unified network using the same training data in the future.

References

[1] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[2] X.-S. Zhang, S.-B. Gao, C.-Y. Li, and Y.-J. Li, "A retina inspired model for enhancing visibility of hazy images," *Frontiers Comput. Neurosci.*, vol. 9, p. 151, Dec. 2015.

[3] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "End-to-end united video dehazing and detection," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 7016–7023

[4] B. Li *et al.* (2017). "Benchmarking single image dehazing and beyond." [Online]. Available: https://arxiv.org/abs/1712.04143

[5] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1956–1963.

[7] K. B. Gibson, D. T. Vo, and T. Q. Nguyen, "An investigation of dehazing effects on image and video coding," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 662–673, Feb. 2012.

[8] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1674–1682.

[9] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 889–896, May 2000.

[10] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Effective single image dehazing by fusion," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3541–3544.

[11] L. Caraffa and J.-P. Tarel, "Markov random field model for single image defogging," in *Proc. IEEE Intell. Vehicles Symp.*, 2013, pp. 994–999.

[12] Y. Wang and C. Fan, "Single image defogging by multiscale depth fusion," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4826–4837, Nov. 2014.

[13] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 13:1–13:14, Dec. 2014.

[14] Z. Li, P. Tan, R. T. Tan, D. Zou, S. Z. Zhou, and L.-F. Cheong, "Simultaneous video defogging and stereo reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4988–4997.

[15] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[17] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3271–3282, Aug. 2013.

[18] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 999–1009, Feb. 2018.

[19] X. Fan, Y. Wang, X. Tang, R. Gao, and Z. Luo, "Two-layer Gaussian process regression with example selection for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2505–2517, Dec. 2017.

[20] J.-H. Kim, W.-D. Jang, Y. Park, D.-H. Lee, J.-Y. Sim, and C.-S. Kim, "Temporally real-time video dehazing," in *Proc. IEEE Int. Conf. Image Process.*, Sep./Oct. 2012, pp. 969–972.

[21] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 237–246.

[22] W. Ren, J. Pan, X. Cao, and M.-H. Yang, "Video deblurring via semantic segmentation and pixel-wise non-linear kernel," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1086–1094.

[23] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.

[24] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–10.

[25] Q. Zhu, J. Mai, and L. Shao, "Single image dehazing using color attenuation prior," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–10.

[26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[27] D. Liu *et al.*, "Robust video super-resolution with learned temporal dynamics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2526–2534.

[28] H. Huang *et al.*, "Real-time neural style transfer for videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7044–7052.

[29] J. Zhang, L. Li, Y. Zhang, G. Yang, X. Cao, and J. Sun, "Video dehazing with spatial and temporal coherence," *Vis. Comput.*, vol. 27, nos. 6–8, pp. 749–757, 2011.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[31] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 162–171.

[32] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7190–7198.

[33] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1339–1348.

[34] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 409–421, 2017.

[35] Z. Shen, W. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8260–8269.

[36] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by spatial feature modulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

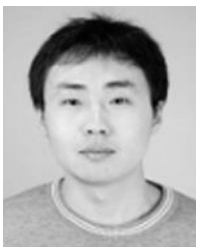
[37] Z. Cheng, S. You, V. Ila, and H. Li. (2018). "Semantic single-image dehazing." [Online]. Available: https://arxiv.org/abs/1804.05624

[38] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

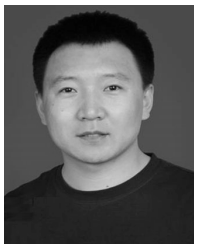
[39] W. Ren and X. Cao, "Deep video dehazing," in *Proc. Pacific-Rim Conf. Multimedia*, 2017, pp. 14–24.

[40] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6978–6986.

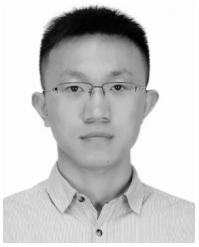
[41] X. Cao, W. Ren, W. Zuo, X. Guo, and H. Foroosh, "Scene text deblurring using text-specific multiscale dictionaries," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1302–1314, Apr. 2015.
[42] R. T. Tan, N. Pettersson, and L. Petersson, "Visibility enhancement for roads with foggy or hazy scenes," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2007, pp. 19–24.
[43] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[44] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.
[45] S. Zhang, W. Ren, and J. Yao, "FEED-Net: Fully end-to-end dehazing," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2018, pp. 1–6.
[46] Y. Yan, W. Ren, and X. Cao, "Recolored image detection via a deep discriminative model," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 5–17, Jan. 2019.
[47] H. Zhang, V. Sindagi, and V. M. Patel. (2017). "Image de-raining using a conditional generative adversarial network." [Online]. Available: https://arxiv.org/abs/1701.05957
[48] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 695–704.
[49] R. Liu, X. Fan, M. Hou, Z. Jiang, Z. Luo, and L. Zhang, "Learning aggregated transmission propagation networks for haze removal and beyond," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[50] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
[51] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
[52] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4770–4778.
[53] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5168–5177.
[54] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 617–624.
[55] C. Chen, M. N. Do, and J. Wang, "Robust image and video dehazing with visual artifact suppression via gradient residual minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 576–591.
[56] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L^1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*. Berlin, Germany: Springer, 2009, pp. 23–45.
[57] M. Delbracio and G. Sapiro, "Hand-held video deblurring via efficient Fourier aggregation," *IEEE Trans. Comput. Imag.*, vol. 1, no. 4, pp. 270–283, Dec. 2015.
[58] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122
[59] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.
[60] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 707–724.
[61] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1685–1694.



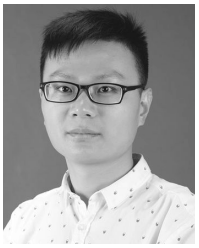
Wenqi Ren received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017. From 2015 to 2016, he was supported by the China Scholarship Council and working with Prof. M.-H. Yang as a joint-training Ph.D. Student with the Electrical Engineering and Computer Science Department, University of California at Merced. He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences, China. His research interests include image processing and related high-level vision problems. He received the Tencent Rhino Bird Elite Graduate Program Scholarship in 2017 and the MSRA Star Track Program in 2018.



Jingang Zhang is currently an Associate Professor with the University of Chinese Academy of Sciences. He has been engaged in the research of computational optical imaging technology for more than 10 years, presided over more than 10 national and ministerial-level scientific research projects, such as the National Natural Science Foundation of China and the Joint Foundation Program of the Chinese Academy of Sciences for equipment pre-feasibility study, more than eight authorized patents and four software copyrights, published over 20 academic papers, and co-authored an academic book *Vector-based Mathematics and Geometr*. His research interests include image denosing, deblurring and dehazing, image/video analysis and enhancement, and related high-level vision problems.



Xiangyu Xu received the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, China, in 2018. He was a joint-training Ph.D. Student with the School of Electrical Engineering and Computer Science, University of California at Merced, Merced, CA, USA, from 2015 to 2017. He is currently a Research Scientist with SenseTime, Beijing. His research interest includes image processing, low-level vision, and deep learning.



Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noah's Ark Laboratory, Hong Kong, from 2013 to 2016. He is currently a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China. His current research interests lie in the areas of computer vision and multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was the Finalist of the HKIS Young Scientist Award in Engineering Science in 2012.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and co-authored more than 120 journal and conference papers. Prof. Cao is a fellow of the IET. He is on the Editorial Board of the IEEE TRANSACTIONS OF IMAGE PROCESSING. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Gaofeng Meng (SM'17) received the B.S. degree in applied mathematics from Northwestern Polytechnical University in 2002, the M.S. degree in applied mathematics from Tianjin University in 2005, and the Ph.D degree in control science and engineering from Xi'an Jiaotong University in 2009. In 2009, he joined the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, as an Assistant Professor, where he was promoted as an Associate Professor in 2012. From 2015 to 2015, he was a Visiting Scholar with the Delft University of Technology, The Netherlands. From 2016 to 2017, he was a Visiting Scholar with Northwestern University, Evanston, IL, USA. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include document image processing, computer vision, and machine learning. He has been serving as an Associate Editor for *Neurocomputing* since 2014.

Wei Liu received the Ph.D. degree in EECS from Columbia University, New York, NY, USA. He was a Research Scientist of the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. He is currently a Distinguished Scientist of Tencent AI Lab and the Director of the Computer Vision Center. He has long been devoted to the Research and Development in the fields of machine learning, computer vision, information retrieval, and big data. Until now, he has published over 180 peer-reviewed journal and conference papers. His research works received a number of awards and honors, such as the 2011 Facebook Fellowship, the 2013 Jury Award for best thesis of Columbia University, the 2016 and 2017 SIGIR Best Paper Award Honorable Mentions, and the 2018 "AI's 10 To Watch" Honor. He also serves as the Area Chair for several international top-tier AI conferences, respectively. He currently serves as an Associate Editor for several international leading AI journals.