The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

End-to-End United Video Dehazing and Detection

Boyi Li,[1]* Xiulian Peng,[2] Zhangyang Wang,[3] Jizheng Xu,[2] Dan Feng[1]
[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[2]Microsoft Research, Beijing, China
[3]Department of Computer Science and Engineering, Texas A&M University
boyilics@gmail.com,xipe@microsoft.com,atlaswang@tamu.edu,jzxu@microsoft.com,dfeng@hust.edu.cn

Abstract

The recent development of CNN-based image dehazing has revealed the effectiveness of end-to-end modeling. However, extending the idea to end-to-end video dehazing has not been explored yet. In this paper, we propose an *End-to-End Video Dehazing Network* (**EVD-Net**), to exploit the temporal consistency between consecutive video frames. A thorough study has been conducted over a number of structure options, to identify the best temporal fusion strategy. Furthermore, we build an *End-to-End United Video Dehazing and Detection Network* (**EVDD-Net**), which concatenates and jointly trains EVD-Net with a video object detection model. The resulting augmented end-to-end pipeline has demonstrated much more stable and accurate detection results in hazy video.

Introduction

The removal of haze from visual data captured in the wild has been attracting tremendous research interests, due to its profound application values in outdoor video surveillance, traffic monitoring and autonomous driving, and so on (Tan 2008). In principle, the generation of hazy visual scene observations follows a known physical model (to be detailed next), and the estimation of key physical parameters, i.e., the atmospheric light magnitude and transmission matrix, become the core step in solving haze removal as an inverse problem (He, Sun, and Tang 2011; Fattal 2014; Berman, Avidan, and others 2016). Recently, the prosperity of convolutional neural networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012) has led to many efforts paid to CNN-based single image dehazing (Ren et al. 2016; Cai et al. 2016; Li et al. 2017a). Among them, DehazeNet (Cai et al. 2016) and MSCNN (Ren et al. 2016) focused on predicting the most important parameter, transmission matrix, from image inputs using CNNs, then generating clean images by the physical model. Lately, AOD-Net (Li et al. 2017a) was the first model to introduce a light-weight end-to-end dehazing convolutional neural network by re-formulating the physical formula. However, there have been only a limited amount of efforts in exploring video dehazing, which is the more realistic scenario, either by traditional statistical approaches or by CNNs.

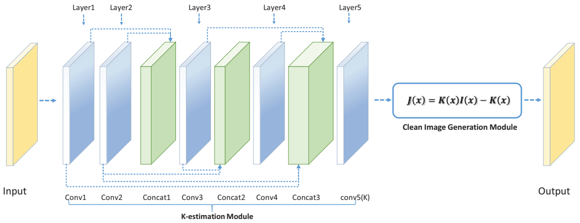


Figure 1: The AOD-Net architecture for single image dehazing (Li et al. 2017a; 2017b), which inspires EVD-Net.

This paper fills in the blank of CNN-based video dehazing by an innovative integration of two important merits in one unified model: (1) we inherit the spirit of training an *end-to-end* model (Li et al. 2017a; Wang et al. 2016), that directly regresses clean images from hazy inputs without any intermediate step. That is proven to outperform the (sub-optimal) results of multi-stage pipelines; (2) we embrace the video setting by explicitly considering how to embed the temporal coherence between neighboring video frames when restoring the current frame. By an extensive architecture study, we identify the most promising temporal fusion strategy, which is both interpretable from a dehazing viewpoint and well aligned with previous findings (Karpathy et al. 2014; Kappeler et al. 2016). We call our proposed model *End-to-End Video Dehazing Network* (**EVD-Net**).

Better yet, EVD-Net can be considered as pre-processing for a subsequent high-level computer vision task, and we can therefore jointly train the concatenated pipeline for the optimized high-level task performance in the presence of haze. Using video object detection as a task example, we build the augmented *End-to-End United Video Dehazing and Detection Network* (**EVDD-Net**), and achieve much more stable and accurate detection results in hazy video.

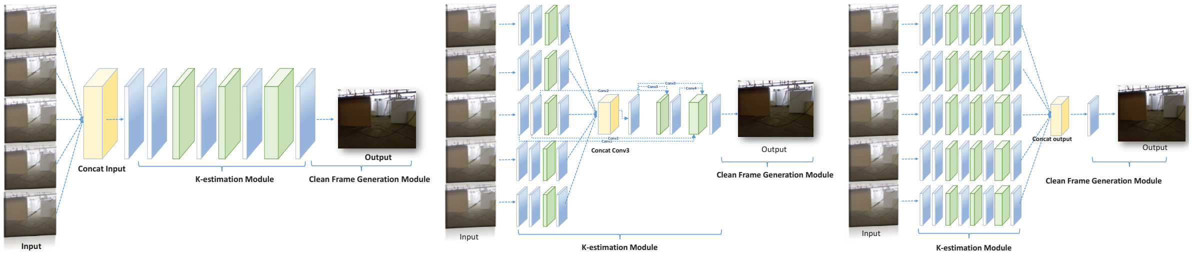
Related Work

Previous single image haze removal algorithms focus on the classical *atmospheric scattering model*:

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

where $I(x)$ is observed hazy image, $J(x)$ is the scene radiance ("clean image") to be recovered. There are two critical

*The work was done at Microsoft Research Asia.



(a) I-level fusion (b) K-level fusion (c) J-level fusion

Figure 2: EVD-Net structure options with 5 consecutive frames as input: (a) I-level fusion, where five input frames are concatenated before feeding the first layer; (b) K-level fusion, where five input frames are first processed separately in its own column and then concatenated after the some layer during K estimation; (c) J-level fusion, where five output images are concatenated.

parameters: A denotes the global atmospheric light, and $t\left(x\right)$ is the transmission matrix defined as:

$$t\left(x\right)=e^{-\beta d\left(x\right)}, \tag{2}$$

where β is the scattering coefficient of the atmosphere, and $d\left(x\right)$ is the distance between the object and the camera. The clean image can thus be obtained in the inverse way:

$$J(x)=\frac{1}{t(x)}I(x)-A\frac{1}{t(x)}+A. \tag{3}$$

A number of methods (Tan 2008; Fattal 2008; He, Sun, and Tang 2011; Meng et al. 2013; Zhu, Mai, and Shao 2015) take advantages of natural image statistics as priors, to predict A and $t\left(x\right)$ separately from the hazy image $I\left(x\right)$. Due to the often inaccurate estimation of either (or both), they tend to bring in many artifacts such as non-smoothness, unnatural color tones or contrasts. Many CNN-based methods (Cai et al. 2016; Ren et al. 2016) employ CNN as a tool to regress $t\left(x\right)$ from $I\left(x\right)$. With A estimated using some other empirical methods, they are then able to estimate $J(x)$ by (3). Notably, (Li et al. 2017a; 2017b) design the first completely end-to-end CNN dehazing model based on re-formulating (1), which directly generates $J(x)$ from $I(x)$ without any other intermediate step:

$$J\left(x\right)=K\left(x\right)I\left(x\right)-K\left(x\right), \text{where}$$
$$K\left(x\right)=\frac{\frac{1}{t(x)}\left(I\left(x\right)-A\right)+A}{I\left(x\right)-1}. \tag{4}$$

Both $\frac{1}{t(x)}$ and A are integrated into the new variable $K\left(x\right)$[1]. As shown in Figure 1, the AOD-Net architecture is composed of two modules: a *K-estimation module* consisting of five convolutional layers to estimate $K\left(x\right)$ from $I\left(x\right)$, followed by a *clean image generation module* to estimate $J\left(x\right)$ from both $K\left(x\right)$ and $I\left(x\right)$ via (4). All those above-mentioned methods are designed for single-image dehazing, without taking into account the temporal dynamics in video.

When it comes to video dehazing, a majority of existing approaches count on post processing to correct temporal inconsistencies, after applying single image dehazing algorithms frame-wise. (Kim et al. 2013) proposes to inject temporal coherence into the cost function, with a clock filter for speed-up. (Li et al. 2015) jointly estimates the scene depth and recovers the clear latent images from a foggy video sequence. (Chen, Do, and Wang 2016) presents an image-guided, depth-edge-aware smoothing algorithm to refine the transmission matrix, and uses Gradient Residual Minimization to recover the haze-free images. (Cai, Xu, and Tao 2016) designs a spatio-temporal optimization for real-time video dehazing. But as our experiments will show, those relatively simple and straightforward video dehazing approaches may not be even able to outperform the sophisticated CNN-based single image dehazing models. The observation reminds us that the utility of temporal coherence must be coupled with more advanced model structures (such as CNNs) for the further boost of video dehazing performance.

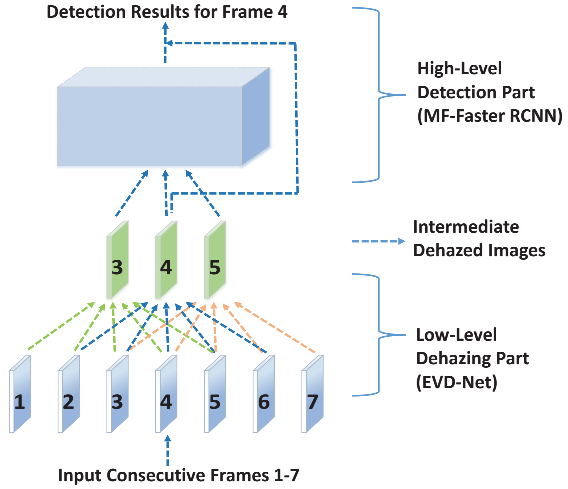


Figure 3: EVDD-Net for united video dehazing and detection, with a tree-like deep architecture. Note that the entire pipeline will be jointly optimized for training.

Recent years have witnessed a growing interest in modeling video using CNNs, for a wide range of tasks such as super-resolution (SR) (Kappeler et al. 2016), deblurring (Su et al. 2016), classification (Karpathy et al. 2014; Shen et al.

[1] There was a constant bias b in (Li et al. 2017a; 2017b), which is omitted here to simplify notations.

2016), and style transfer (Chen et al. 2017). (Kappeler et al. 2016) investigates a variety of structure configurations for video SR. Similar attempts are made by (Karpathy et al. 2014; Shen et al. 2016), both digging into different connectivity options for video classification. (Liu et al. 2017) proposes a more flexible formulation by placing a spatial alignment network between frames.(Su et al. 2016) introduces a CNN trained end-to-end to learn accumulating information across frames for video deblurring. For video style transfer, (Chen et al. 2017) incorporates both short-term and long-term coherences and also indicates the superiority of multi-frame methods over single-frame ones.

End-to-End Video Dehazing Network and Its Unity with Video Detection

We choose the AOD-Net model (Li et al. 2017a; 2017b) for single image dehazing as the starting point to develop our deep video dehazing model, while recognizing that the proposed methodology can be applied to extending other deep image dehazing models to video, e.g., (Cai et al. 2016; Ren et al. 2016). Our main problem lies in the strategy of *temporal fusion*. As a well-justified fact in video processing, jointly considering neighboring frames when predicting the current frame will benefit many image restoration and classification-type tasks (Liu and Sun 2014; Ma et al. 2015; Kappeler et al. 2016). Specifically to the video dehazing case, both object depth (which decides the transmission matrix T) and the global atmospheric light A should be hardly or slowly changed over a moderate number of consecutive frames, implying the great promise of exploiting multi-frame coherence for video dehazing.

Fusion Strategy for Video Dehazing: Three Structure Options

Enlightened by the analysis from (Kappeler et al. 2016), we investigate three different strategies to fuse consecutive frames. For simplicity, we show the architecture for five input frames as an example, namely the previous two $(t-2, t-1)$, current (t), and next two $(t+1, t+2)$ frames, with the goal to predict the clean version for the current frame t. Clearly, any number of past and future frames can be accommodated. As compared in Figure 2, three different types of fusion structures are available for EVD-Net:

- I**-Level Fusion:** fusing at the input level. All five input frames are concatenated along the first dimension before the first convolutional layer is applied. It corresponds to directly fusing image features at the pixel level, and then running single-image dehazing model on the fused image.
- K**-Level Fusion:** fusing during the K-estimation. Specifically, we will term the following structure as *K-level fusion, conv l* (l = 1, 2, ... 5): each input frame will go through the first l convolutional layers separately before concatenation at the output of the l-th layer, l = 1, 2, ... 5. In other words, the multi-frame information is fused towards generating the key parameter K (i.e., $t(x)$ and A) of the current frame, based on the underlying assumption that both object depths and global atmospheric light transmit smoothly across neighboring frames.
- J**-Level Fusion:** fusing during the output level. It is equivalent to feed each frame to its separate K-estimation module, and the five K outputs are concatenated right before the clean image generation module. It will not fuse until all frame-wise predictions have been made, and corresponds to fusing at the output level.

Training a video-based deep model is often more hassle. (Kappeler et al. 2016) proves that a well-trained single-column deep model for images could provide a high-quality initialization for training a multi-column model for videos, by splitting all convolutional weights before the fusion step. We follow their strategy, training an AOD-Net first to initialize different EVD-Net architectures in EVD-Net.

Unity Brings Power: Optimizing Dehazing and Detection as An End-to-End Pipeline in Video

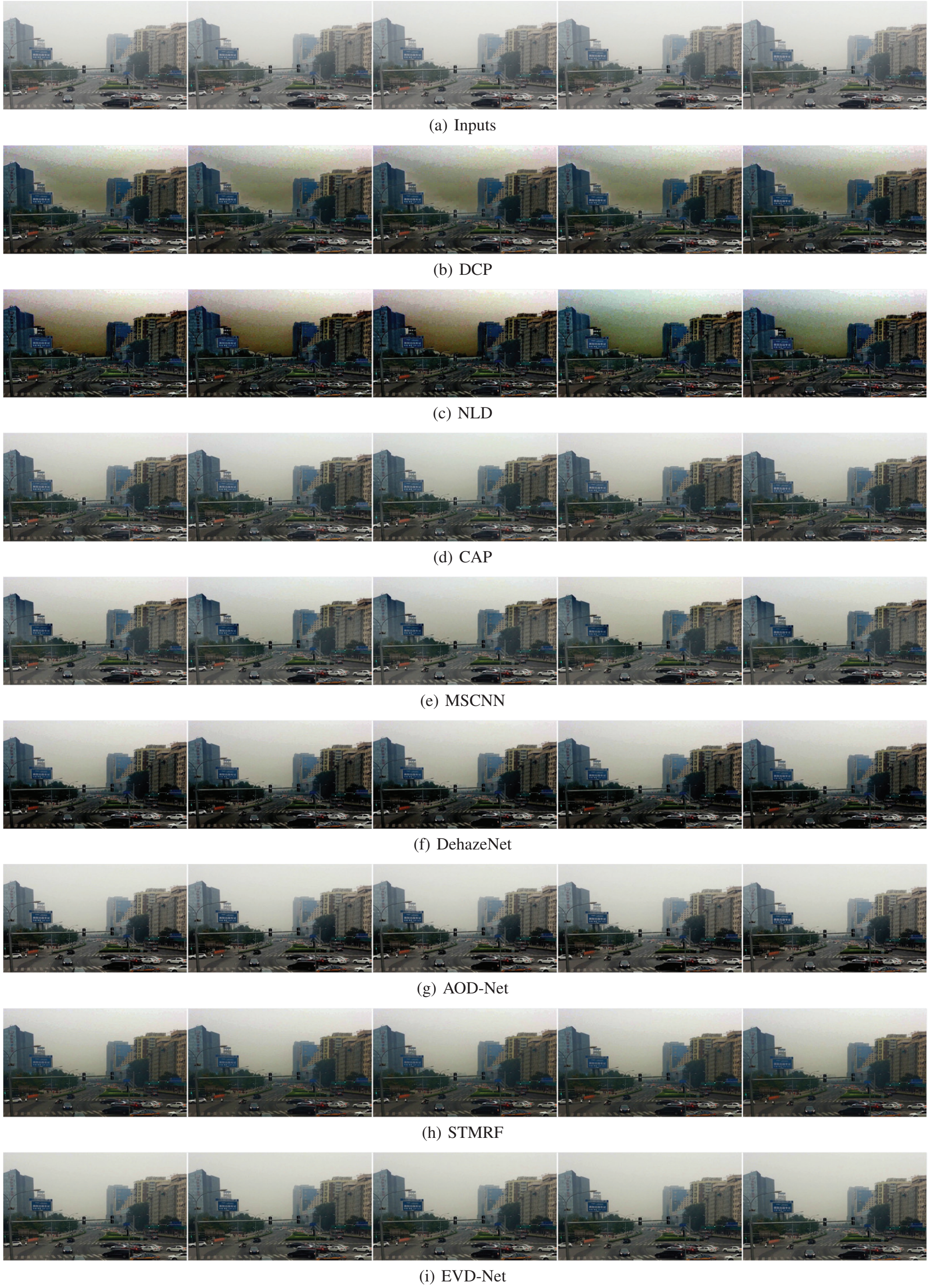
Beyond the video restoration purpose, dehazing, same as many other low-level restoration and enhancement techniques, is commonly employed as pre-processing, to improve the performance of high-level computer vision tasks in the presence of certain visual data degradations. A few pioneering works in single-image cases (Wang et al. 2016; Li et al. 2017a; 2017b) have demonstrated that formulating the low-level and high-level tasks(Ren et al. 2015; Wang et al. 2012; 2017) as one unified (deep) pipeline and optimizing it from end to end will convincingly boost the performance. Up to our best knowledge, the methodology has not been validated in video cases yet.

In outdoor surveillance or autonomous driving, object detection from video (Kang et al. 2016; Tripathi et al. 2016; Zhu et al. 2017) is widely desirable, whose performance is known to heavily suffer from the existence of haze. For example, autonomous vehicles rely on a light detection and ranging (LIDAR) sensor to model the surrounding world, and a video camera (and computer, mounted in the vehicle) records, analyzes and interprets objects visually to create 3D maps. However, haze can interfere with laser light from the LIDAR sensor and fail subsequent algorithms.

In this paper, we investigate the brand-new joint optimization pipeline of video dehazing and video object detection. Beyond the dehazing part, the detection part has to take into account temporal coherence as well, to reduce the flickering detection results. With EVD-Net, we further design a video-adapted version of Faster R-CNN (Ren et al. 2015) and verify its effectiveness, while again recognizing the possibility of plugging in other video detection models. For the first two convolutional layers in the classical single-image Faster R-CNN model, we split them into three parallel branches to input the previous, current, and next frames, respectively[2]. They are concatenated after the second convolutional layer, and go through the remaining layers to predict object bounding boxes for the current frame. We call it *Multi-Frame Faster R-CNN* (**MF-Faster R-CNN**).

Finally, uniting EVD-Net and MF-Faster R-CNN in one gives rise to EVDD-Net, which naturally displays an interesting locally-connected, tree-like structure and is subject to further (and crucial) joint optimization. Figure 3 plots an

[2] The window size 3 here is by default, but could be adjusted.



(a) Inputs

(b) DCP

(c) NLD

(d) CAP

(e) MSCNN

(f) DehazeNet

(g) AOD-Net

(h) STMRF

(i) EVD-Net

Figure 4: Challenging natural consecutive frames results compared with the state-of-art methods.

instance of EVDD-Net, with a *low-level temporal window size* of 5 frames, and a *high-level temporal window size* of 3 frames, leading to the *overall temporal window size* of 7 frames. We first feed 7 consecutive frames (indexed at 1, 2, ..., 7) into the EVDD-Net part. By predicting on 5-frame groups with a stride size of 1, three dehazed results corresponding to the frames 3, 4, 5 will be generated. They are then fed into the MF-Faster R-CNN part to fuse the detection results of frame 4. Essentially, the tree-like structure comes from the two-step utilization of temporal coherence between

Table 1: PSNR/SSIM Comparisons of Various Structures.

Methods	PSNR	SSIM
I-level fusion, 3 frames	20.5551	0.8515
I-level fusion, 5 frames	20.6095	0.8529
K-level fusion, conv1, 3 frames	20.6105	0.9076
K-level fusion, conv1, 5 frames	20.8240	0.9107
K-level fusion, conv2, 3 frames	20.6998	0.9028
K-level fusion, conv2, 5 frames	**20.9908**	**0.9087**
K-level fusion, conv2, 7 frames	20.7901	0.9049
K-level fusion, conv2, 9 frames	20.7355	0.9042
K-level fusion, conv3, 3 frames	20.9187	0.9078
K-level fusion, conv3, 5 frames	20.7780	0.9051
K-level fusion, conv4, 3 frames	20.7468	0.9038
K-level fusion, conv4, 5 frames	20.6756	0.9027
K-level fusion, conv5(K), 3 frames	20.6546	0.8999
K-level fusion, conv5(K), 5 frames	20.7942	0.9046
J-level fusion, 3 frames	20.4116	0.8812
J-level fusion, 5 frames	20.3675	0.8791

Table 2: PSNR/SSIM Comparisons of Various Approaches.

Methods	PSNR	SSIM
ATM (Sulami et al. 2014)	11.4190	0.6534
BCCR (Meng et al. 2013)	13.4206	0.7068
NLD (Berman, Avidan, and others 2016)	13.9059	0.6456
FVR (Tarel and Hautiere 2009)	16.2945	0.7799
DCP (He, Sun, and Tang 2011)	16.4499	0.8188
DehazeNet (Cai et al. 2016)	17.9332	0.7963
CAP (Zhu, Mai, and Shao 2015)	20.4097	0.8848
MSCNN (Ren et al. 2016)	20.4839	0.8690
AOD-Net (Li et al. 2017a)	20.6828	0.8549
STMRF (Cai, Xu, and Tao 2016)	18.9956	0.8707
EVD-Net	**20.9908**	**0.9087**

neighboring frames, in both low level and high level. We are confident that such a tree-like structure will be of extensive reference values to more future deep pipelines that seek to jointly optimize low-level and high-level tasks.

Experiment Results on Video Dehazing

Datasets and Implementation We created a synthetic hazy video dataset based on (1), using 18 videos selected from the TUM RGB-D Dataset (Sturm et al. 2012), which captures varied visual scenes. The depth information is refined by the filling algorithm in (Silberman et al. 2012). We then split it into a training set, consisting of 5 videos with 100,000 frames, and a non-overlapping testing set called *TestSet V1*, consisting of the rest 13 relatively short video clips with a total of 403 frames.

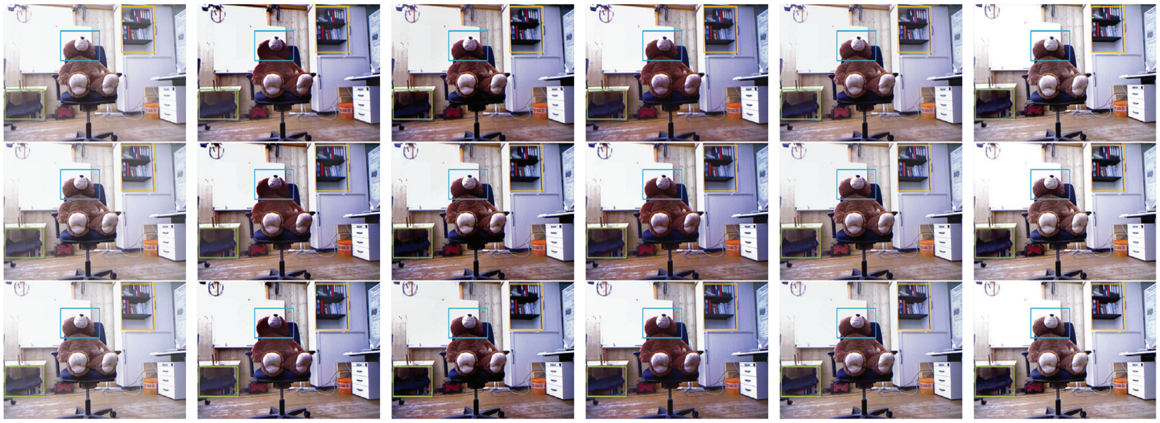
When training EVD-Net, the momentum and the decay parameters are set to 0.9 and 0.0001, respectively, with a batch size of 8. We adopt the Mean Square Error (MSE) loss, which has been shown in (Li et al. 2017a; 2017b) that it is well aligned with SSIM and visual quality. Thanks to the light-weight structure, EVD-Net takes only 8 epochs (100,000 iterations) to converge.

Fusion Structure Comparison We first compare the performances of three fusion strategies in EVD-Net with different configuration parameters on TestSet V1. As shown in Table 1, the performance of K-level fusion is far superior to I-level fusion and J-level fusion in both PSNR and SSIM, albeit the number of network parameters in J-level fusion is much more than the other two. Moreover, among all configurations of K-level fusion, when using 3 input frames, *K-level fusion, conv 3* performs the best. While using 5 input frames the performance further increases and reaches an overall peak at *K-level fusion, conv 2*, and is chosen as the default configuration of EVD-Net. When testing more frames such as 7 or 9, we observe the performance gets saturated and sometimes hurt, since the relevance of far-away frames to the current frame will decay fast.

Quantitative Comparison We compare EVD-Net on TestSet V1 with a variety of state-of-the-art single image dehazing methods, including: Automatic Atmospheric Light Recovery (**ATM**) (Sulami et al. 2014), Boundary Constrained Context Regularization (**BCCR**) (Meng et al. 2013), Fast Visibility Restoration (**FVR**) (Tarel and Hautiere 2009), Non-local Image Dehazing (**NLD**) (Berman, Avidan, and others 2016; Berman, Treibitz, and Avidan 2017), Dark-Channel Prior (**DCP**) (He, Sun, and Tang 2011), **MSCNN** (Ren et al. 2016), **DehazeNet** (Cai et al. 2016), Color Attenuation Prior (**CAP**) (Zhu, Mai, and Shao 2015), and **AOD-Net** (Li et al. 2017a). We also compare with a recently proposed video dehazing approach: Real-time Dehazing Based on Spatio-temporal MRF (**STMRF**) (Cai, Xu, and Tao 2016). Table 2 demonstrates the very promising performance margin of EVD-Net over others, in terms of both PSNR and SSIM. Compared to the second best approach of AOD-Net, EVD-Net gains an advantage of over 0.3 dB in PSNR and 0.05 in SSIM, showing the benefits of temporal coherence. Compared to the video-based STMRF (which is not CNN-based), we notice a remarkable performance gap of 2 dB in PSNR and 0.04 in SSIM.

Qualitative Visual Quality Comparision Figure 4 shows the comparison results on five consecutive frames for a number of image and video dehazing approaches, over a real-world hazy video (with no clean ground-truth). The test video is taken from a city road when the PM 2.5 is 223, constituting a challenging heavy haze scenario. Without the aid of temporal coherence, single image dehazing approaches tend to produce temporal inconsistencies and jaggy artifacts. The DCP and NLD results are especially visually unpleasing. CAP and MSCNN, as well as STMRF, fail to fully remove haze, e.g., in some building areas (please amplify to view details), while DehazeNet tends to darken the global light. AOD-Net produces reasonably good results, but sometimes cannot ensure the temporal consistencies of illumination and color tones. EVD-Net gives rise to the most visually pleasing, detail-preserving and temporally consistent dehazed results among all.

Figure 5 shows a comparison example on synthetic data, where three consecutive frames are selected from TestSet V1. By comparing to the ground-truth, it can be seen that EVD-Net again preserves both details and color tones best.



(a) Inputs (b) MSCNN (c) DehazeNet (d) STMRF (e) EVD-Net (f) Groundtruth

Figure 5: Synthetic consecutive frames results compared with the state-of-art methods and Groundtruth frames.

Table 3: Average Precision(AP) of each categories and Mean Average Precision (MAP) on TestSet V2.

Metrics	Original Faster R-CNN	Re-trained Faster R-CNN	EVD+Faster R-CNN	JAOD-Faster R-CNN	**EVDD-Net**
Car AP	0.810	0.807	0.811	0.808	0.803
Bicycle AP	0.531	0.703	0.603	0.707	0.802
MAP	0.671	0.755	0.707	0.758	**0.802**

Experiment Results on Video Detection

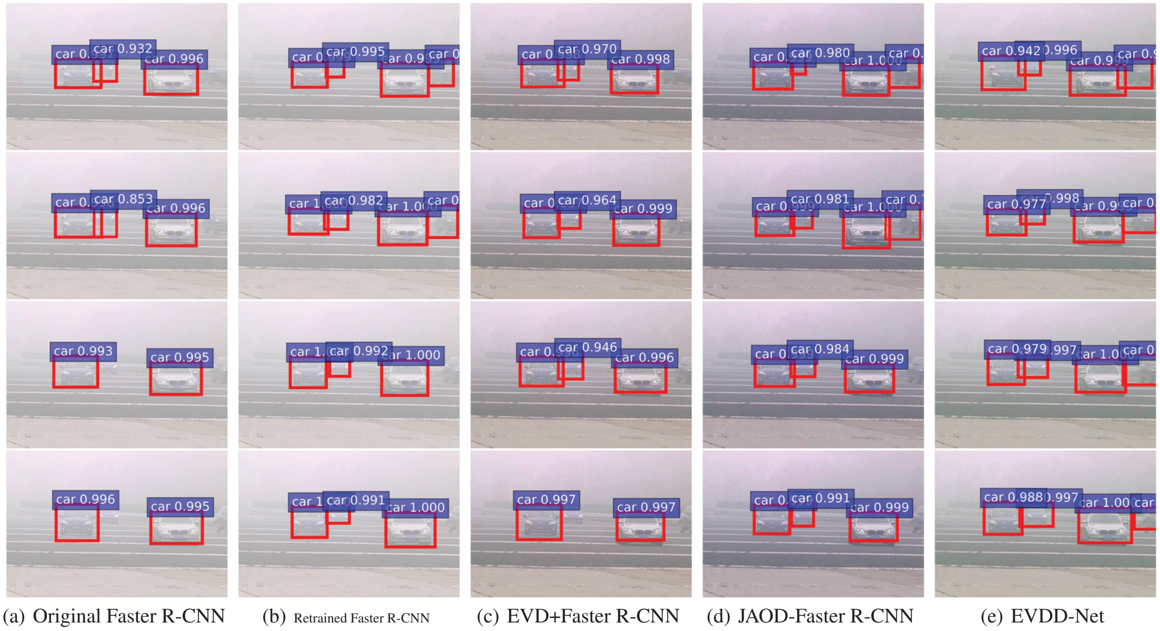
Datasets and Implementation While training EVDD-Net, the lack of hazy video datasets with object detection labels has again driven us to create our own synthetic training set. We synthesize hazy videos with various haze levels for a subset of ILSVRC2015 VID dataset (Russakovsky et al. 2015) based on the atmospheric scattering model in (1) and estimated depth using the method in (Liu et al. 2016). The EVDD-Net is trained using 4,499 frames from 48 hazy videos for a two-category object detection problem (car, bike), and tested on 1,634 frames from another 10 hazy videos (*TestSet V2*). Several real-world hazy videos are also used for evaluation.

The training of EVDD-Net evidently benefits from high-quality initialization: a trained EVD-Net, plus a MF-Faster RCNN model initialized by splitting the first two convolutional layers similar to the way in (Kappeler et al. 2016). While (Li et al. 2017b) found that directly end-to-end training of two parts could lead to sufficiently good results, we observe that the video-based pipeline involves much more parameters and are thus more difficult to train end to end. Besides the initialization, we also find a two-step training strategy for EVDD-Net: we first tune only the fully-connected layers in the high-level detection part of EVD-Net for 90,000 iterations, and then tune the entire concatenated pipeline for another 10,000 iterations.

Comparison Baselines EVDD-Net is compared against a few baselines: i) the *original Faster R-CNN* (Ren et al. 2015), which is single image-based and trained on haze-free images; ii) *Re-trained Faster R-CNN*, which is obtained by retraining the original Faster R-CNN on a hazy image dataset; iii) *EVD + Faster R-CNN*, which is a simple concatenation of separately trained EVD-Net and original Faster R-CNN models; iv) *JAOD-Faster R-CNN*, which is the state-of-the-art single-image joint dehazing and detection pipeline proposed in (Li et al. 2017b).

Results and Analysis Table 3 presents the Mean Average Precision (MAP) of all five approaches, which is our main evaluation criterion. We also display the category-wise average precision for references. Comparing the first two columns verify that the object detection algorithms trained on conventional visual data do not generalize well on hazy data. Directly placing EVD-Net in front of MF-Faster R-CNN fails to outperform Retrained Faster-RCNN, although it surpasses the original Faster-RCNN with a margin. We notice that it coincides with some earlier observations in other degradation contexts (Wang et al. 2016), that a naive concatenation of low-level and high-level models often cannot sufficiently boost the high-level task performance, as the low-level model will simultaneously bring in recognizable details and artifacts. The performance of JAOD-Faster R-CNN is promising, and slightly outperforms Retrained Faster-RCNN. However, its results often show temporally flickering and inconsistent detections. EVDD-Net achieves a significantly boosted MAP over other baselines. EVDD-Net is another successful example of "closing the loop" of low-level and high-level tasks, based on the well-verified assumption that the degraded image, if correctly restored, will also have a good identifiability.

Figure 6 shows a group of consecutive frames and object detection results for each approach, from a real-world hazy video sequence. EVDD-Net is able to produce both the most accurate and temporally consistent detection results. In this specific scene, EVDD-Net is the only approach that can correctly detect all four cars throughout the four displayed



(a) Original Faster R-CNN (b) Retrained Faster R-CNN (c) EVD+Faster R-CNN (d) JAOD-Faster R-CNN (e) EVDD-Net

Figure 6: Comparisons of detection results on real-world hazy video sample frames. Note that for the third, fourth and fifth columns, the results are visualized on top of the (intermediate) dehazing results.

frames, especially the rightmost car that is hardly recognizable even for human eyes. That is owing to the temporal regularizations in both low-level and high-level parts of EVDD-Net. More video results can be found in the YouTube[3].

Conclusion

This paper proposes EVD-Net, the first CNN-based, fully end-to-end video dehazing model, and thoroughly investigates the fusion strategies. Furthermore, EVD-Net is concatenated and jointly trained with a video object detection model, to constitute an end-to-end pipeline called EVDD-Net, for detecting objects in hazy video. Both EVD-Net and EVDD-Net are extensively evaluated on synthetic and real-world datasets, to verifyt the dramatic superiority in both dehazing quality and detection accuracy. Our future work aims to strengthen the video detection part of EVDD-Net.

Acknowledgement

Boyi Li and Dan Feng[4]'s research are in part supported by the National Natural Science Foundation of China No.61772222, Shenzhen Research Funding of Science and Technology JCYJ20170307172447622.

References

Berman, D.; Avidan, S.; et al. 2016. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1674–1682.

Berman, D.; Treibitz, T.; and Avidan, S. 2017. Air-light estimation using haze-lines. In *Computational Photography (ICCP), 2017 IEEE International Conference on*, 1–9.

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing* 25(11).

Cai, B.; Xu, X.; and Tao, D. 2016. Real-time video dehazing based on spatio-temporal mrf. In *Pacific Rim Conference on Multimedia*, 315–325. Springer.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017. Coherent online video style transfer. *arXiv preprint arXiv:1703.09211*.

Chen, C.; Do, M. N.; and Wang, J. 2016. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *European Conference on Computer Vision*, 576–591.

Fattal, R. 2008. Single image dehazing. *ACM transactions on graphics (TOG)* 27(3):72.

Fattal, R. 2014. Dehazing using color-lines. *ACM Transactions on Graphics (TOG)* 34(1):13.

He, K.; Sun, J.; and Tang, X. 2011. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* 33(12):2341–2353.

Kang, K.; Ouyang, W.; Li, H.; and Wang, X. 2016. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 817–825.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging* 2(2):109–122.

[3] https://youtu.be/Lih7Q91ykUk

[4] Dan Feng is the corresponding author of this paper.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kim, J.-H.; Jang, W.-D.; Sim, J.-Y.; and Kim, C.-S. 2013. Optimized contrast enhancement for real-time image and video dehazing. *Journal of Visual Communication and Image Representation* 24(3):410–425.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, Z.; Tan, P.; Tan, R. T.; Zou, D.; Zhiying Zhou, S.; and Cheong, L.-F. 2015. Simultaneous video defogging and stereo reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4988–4997.

Li, B.; Peng, X.; Wang, Z.; Xu, J.-Z.; and Feng, D. 2017a. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*.

Li, B.; Peng, X.; Wang, Z.; Xu, J.; and Feng, D. 2017b. An all-in-one network for dehazing and beyond. *arXiv preprint arXiv:1707.06543*.

Liu, C., and Sun, D. 2014. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence* 36(2):346–360.

Liu, F.; Shen, C.; Lin, G.; and Reid, I. 2016. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence* 38(10):2024–2039.

Liu, D.; Wang, Z.; Fan, Y.; Liu, X.; Wang, Z.; Chang, S.; and Huang, T. S. 2017. Robust video super-resolution with learned temporal dynamics. In *IEEE International Conference on Computer Vision*.

Ma, Z.; Liao, R.; Tao, X.; Xu, L.; Jia, J.; and Wu, E. 2015. Handling motion blur in multi-frame super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5224–5232.

Meng, G.; Wang, Y.; Duan, J.; Xiang, S.; and Pan, C. 2013. Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision*, 617–624.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 91–99.

Ren, W.; Liu, S.; Zhang, H.; Pan, J.; Cao, X.; and Yang, M.-H. 2016. Single image dehazing via multi-scale convolutional neural networks. In *European Conference on Computer Vision*, 154–169.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Shen, H.; Han, S.; Philipose, M.; and Krishnamurthy, A. 2016. Fast video classification via adaptive cascading of deep models. *arXiv preprint arXiv:1611.06453*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Nyu depth dataset v2. ECCV.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems*.

Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; and Wang, O. 2016. Deep video deblurring. *arXiv preprint arXiv:1611.08387*.

Sulami, M.; Glatzer, I.; Fattal, R.; and Werman, M. 2014. Automatic recovery of the atmospheric light in hazy images. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, 1–11. IEEE.

Tan, R. T. 2008. Visibility in bad weather from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Tarel, J.-P., and Hautiere, N. 2009. Fast visibility restoration from a single color or gray level image. In *Computer Vision, IEEE 12th International Conference on*, 2201–2208.

Tripathi, S.; Lipton, Z. C.; Belongie, S.; and Nguyen, T. 2016. Context matters: Refining object detection in video with recurrent neural networks. *arXiv preprint arXiv:1607.04648*.

Wang, M.; Hong, R.; Li, G.; Zha, Z.-J.; Yan, S.; and Chua, T.-S. 2012. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia* 14(4):975–985.

Wang, Z.; Chang, S.; Yang, Y.; Liu, D.; and Huang, T. S. 2016. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4792–4800.

Wang, M.; Luo, C.; Ni, B.; Yuan, J.; Wang, J.; and Yan, S. 2017. First-person daily activity recognition with manipulated object proposals and non-linear feature fusion. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Flow-guided feature aggregation for video object detection. *arXiv preprint arXiv:1703.10025*.

Zhu, Q.; Mai, J.; and Shao, L. 2015. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing* 24(11):3522–3533.