Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Study

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin

***Abstract*— Existing enhancement methods are empirically expected to help the high-level end computer vision task: however, that is observed to not always be the case in practice. We focus on object or face detection in poor visibility enhancements caused by bad weathers (haze, rain) and low light conditions. To provide a more thorough examination and fair comparison, we introduce three benchmark sets collected in real-world hazy, rainy, and low-light conditions, respectively, with annotated objects/faces. We launched the UG^{2+} challenge Track 2 competition in IEEE CVPR 2019, aiming to evoke a comprehensive discussion and exploration about whether and how low-level vision techniques can benefit the high-level automatic visual recognition in various scenarios. To our best knowledge, this is the first and currently largest effort of its kind. Baseline results by cascading existing enhancement and detection models are reported, indicating the highly challenging nature of our new data as well as the large room for further technical innovations. Thanks to a large participation from the research community, we are able to analyze representative team solutions, striving to better identify the strengths and limitations of existing mindsets as well as the future directions.**

***Index Terms*— Poor visibility environment, object detection, face detection, haze, rain, low-light conditions.**

I. Introduction

THE arrival of the big data era brings us mass diverse applications and spawns a series of demands in both human and machine visions. On one hand, new applications in consumer electronics [1], such as TV broadcasting, movies, video-on-demand, *etc*., expect continuous efforts to improve the human visual experience. On the other hand, many applications of smart cities and Internet of things (IoT), such as surveillance video, autonomous/assisted driving, unmanned aerial vehicles (UAVs), *etc.*, call for more effective and stable machine vision-based sensing and understanding [2]. As a result, it is a critical issue to explore the general framework that can benefit two kinds of tasks simultaneously and make them mutually beneficial.

However, most of the existing researches still aim to solve the problems in their own routes separately. In early researches [3]–[5], their models are not capable enough to consider beyond their own purposes (only for one of human vision or machine vision). Even in the recent decade, most of the enhancement methods, *e.g.*, dehazing [6]–[9], deraining [10]–[19], illumination enhancement [20]–[25], image/video compression [26]–[28], compression artifact removal [29], [30], and copy-move forgery detection [31], only target human vision and most of image/video understanding and analytics methods, *e.g.* classification [32], segmentation [33], action recognition [34], and human pose estimation [35], are considered to take the clean and high-quality images as the input of a system.

When the improved computation power and the new emerging data-driven approaches push for the progress of applying the existing state-of-the-art methods to industrial trials, lack of consideration on human and machine vision jointly leads to the observed fragility of real systems. Taking autonomous driving as an example: the industry players have been tackling the challenges posed by inclement weathers; However, heavy rain, haze or snow will still obscure the vision of on-board cameras and create confusing reflections and glare, leaving the state-of-the-art self-driving cars in the struggle [36]. Another illustrative example can be found in city surveillance: even the commercialized cameras adopted by governments appear fragile in challenging weather conditions [37].

The largely jeopardized performance of visual sensing is caused by two aspects: *inconsistency between training and testing data*, *inappropriate guidance for network training*. First, most current vision systems are designed to perform in clear environments but the real-world scenes are unconstrained

Manuscript received September 2, 2019; revised January 23, 2020 and February 16, 2020; accepted February 16, 2020. Date of publication March 27, 2020; date of current version April 22, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiwen Lu. *(Wenhan Yang and Ye Yuan contributed equally to this work.) (Corresponding authors: Walter J. Scheirer; Zhangyang Wang.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/TIP.2020.2981922

and might include dynamic degradation, *e.g.* moving platforms, bad weathers, and poor illumination, which are not covered in the training data for pretrained models. Second, the existing data-driven methods largely rely on task-driven tuning and extract task-related features. Therefore, they are sensitive to unseen contents and conditions. Once the model faces different degradation to the trained one or has a different target, *e.g.* switching from human vision to machine vision, the performance of the pretrained model might degrade much.

To face the real world in practical applications, a dependable vision system must reckon with the entire spectrum of complex unconstrained outdoor environments, instead of just working in limited scenes. However, at this point, existing academia and industrial solutions see significant gaps from addressing the above-mentioned pressing real-world challenges, and a systematic consideration and collective effort for identifying and resolving those bottlenecks that they commonly face have also been absent. Considering that, it is highly desirable to study to what extent, and in what sense, such challenging visual conditions can be coped with, for the goal of achieving robust visual sensing and understanding in the wild, which benefits security/safety, autonomous driving, robotics, and an even broader range of signal/image processing applications.

One primary challenge arises from the **Data** aspect. Those challenging visual conditions usually give rise to nonlinear and data-dependent degradations that will be much more complicated than the well-studied noise or motion blur. The state-of-the-art deep learning methods are typically hungry for training data. The usage of synthetic training data has been prevailing, but may inevitably lead to domain shifts [38]. Fortunately, those degradations often follow some parameterized physical models. That will naturally motivate a combination of model-based and data-driven approaches. In addition to training, the lack of real-world test sets (and consequently, the usage of potentially oversimplified synthetic sets) has limited the practical scope of the developed algorithms.

The other main challenge is found in the **Goal** side. Most restoration or enhancement methods cast the handling of those challenging conditions as a post-processing step of signal restoration or enhancement after sensing, and then feed the restored data for visual understanding. The performance of high-level visual understanding tasks will thus largely depend on the quality of restoration or enhancement. Yet it remains questionable whether restoration-based approaches would actually boost the visual understanding performance, as the restoration/enhancement step is not optimized towards the target task and may bring in misleading information and artifacts too. For example, a recent line of researches [8], [39]–[48] discuss on the intrinsic interplay relationship of low-level vision and high-level recognition/detection tasks, showing that their goals are not always aligned.

UG^{2+} Challenge Track 2 aims to evaluate and advance the robustness of object detection algorithms in specific poor-visibility situations, including challenging weather and lighting conditions. We structure Challenge 2 into three sub-challenges. Each challenge features a different poor-visibility outdoor condition, and diverse training protocols (paired versus unpaired images, annotated versus unannotated, *etc.*). For each sub-challenge, we collect a new benchmark dataset captured in realistic poor-visibility environments with real image artifacts caused by rain, haze, insufficiency of light. The specific dataset details and evaluation protocols are illustrated in Section III. Comparing with previous works and challenges, our challenge and datasets include the following new features:

- *Covering Complex Degradation:* Our datasets capture images with synthetic and real haze (Challenge 2.1), under-exposure (Challenge 2.2), and rain streaks and raindrops (Challenge 2.3), which provides precious resources to measure the statistics and properties of captured images in these scenes and design effective methods to recover the clean version of these images.
- *Supporting Un/Semi/Full-Supervised Learning:* In Challenge 2.1 and 2.2, our proposed datasets include paired and unpaired data to support both full-supervised or semi-supervised (optional for participants) training. In Challenge 2.3, there is no training data and testing samples provided. Therefore, it is a zero-shot problem and close to unsupervised learning. Therefore, our challenge supports all kinds of learning and provides important materials for future researches.
- *Highly Challenging:* In Challenge 2.1 and 2.2, the winners only achieve the results below 65 MAP. In Challenge 2.3, no participates achieve the results superior to the baseline results. The results show that, our datasets are challenging and there is still large room for further improvement.
- *Full Purpose:* Three sub-challenges support to evaluate the high-level tasks for machine vision. The Challenge 2.1 and 2.2 also include the paired data, which can support to evaluate for human vision.

The rest of the article is organized as follows. Section II briefly reviews previous works on poor visibility enhancement and visual recognition under adverse conditions as well as the related dataset efforts. Section III provides the detailed introduction of our datasets, challenges, evaluation protocols and baseline results. Section V illustrates the competition results and related analysis. Section VI summarizes the interesting observations, the reflected insights and briefly discusses the future directions. The concluding remarks are provided in Section VII.

II. Related Work

A. Datasets

Most datasets used for image enhancement/processing mainly targets at evaluating the quantitative (PSNR, SSIM, *etc.*) or qualitative (visual subjective quality) differences of enhanced images *w.r.t.* the ground truths. Some earlier classical datasets include Set5 [49], Set14 [50], and LIVE1 [51]. The numbers of their images are small. Subsequent datasets come with more diverse scene content, such as BSD500 [52] and Urban100 [53]. The popularity of deep learning methods has increased demand for training and testing data. Therefore, many newer and larger datasets are presented for image and video restoration, such as DIV2K [54] and MANGA109 [55] for image super-resolution, PolyU [56] and Darmstadt [57] for

denoising, RawInDark [58] and LOL dataset [59] for low light enhancement, HazeRD [60], OHAZE [61] and IHAZE [62] for dehazing, Rain100L/H [17] and Rain800 [63] for rain streak removal, and RAINDROP [19] for raindrop removal. However, these datasets provide no integration with subsequent high-level tasks.

A few works [64]–[66] make preliminary attempts for event/action understanding, video summarization, or face recognition in unconstrained and potentially degraded environments. The following datasets are collected by aerial vehicles, including VIRAT Video Dataset [67] for event recognition, UAV123 [68] for UAV tracking, and a multi-purpose dataset [69]. In [70], an unconstrained Face Detection Dataset (UFDD) is proposed for face detection in adverse condition including weather-based degradations, motion blur, focus blur and several others, containing a total of 6,425 images with 10,897 face annotations. However, few works specifically consider the impacts of image enhancement and object detection/recognition jointly. Prior to this UG^{2+} effort, a number of latest works have taken the first stabs. A large-scale hazy image dataset and a comprehensive study – REalistic Single Image DEhazing (RESIDE) [71] – including paired synthetic data and unpaired real data is proposed to thoroughly examine visual reconstruction and vision recognition in hazy images. In [72], an Exclusively Dark (ExDARK) dataset is proposed with a collection of 7,363 images captured from very low-light environments with 12 object classes annotated on both image class level and local object bounding boxes. In [73], the authors present a new large-scale benchmark called RESIDE and a comprehensive study and evaluation of existing single image deraining algorithms, ranging from full-reference metrics, to no-reference metrics, to subjective evaluation and the novel task-driven evaluation. Those datasets and studies shed new light on the comparisons and limitations of state-of-the-art algorithms, and suggest promising future directions. In this work, we follow the footsteps of predecessors to advance the fields by proposing new benchmarks.

B. Poor Visibility Enhancement

There are numerous algorithms aiming to enhance visibility of the degraded imagery, such as image and video denoising/inpainting [74]–[78], deblurring [79], [80], [159], [163], [164] super-resolution [81]–[84] and interpolation [85]. Here we focus on dehazing, low-light condition, and deraining, as in the UG^{2+} Track 2 scope.

1) Dehazing: Dehazing methods proposed in an early stage rely on the exploitation of natural image priors and depth statistics, *e.g.* locally constant constraints and decorrelation of the transmission [86], dark channel prior [9], color attenuation prior [87], nonlocal prior [88]. In [89], [90], Retinex theory is utilized to approximate the spectral properties of object surfaces by the ratio of the reflected light. Recently, Convolutional Neural Network (CNN)-based methods bring in the new prosperity for dehazing. Several methods [6], [7] rely on various CNNs to learn the transmission fully from data. Beyond estimating the haze related variables separately, successive works make their efforts to estimate them in a unified way. In [91], [92], the authors use a factorial Markov random field that integrates the estimation of transmission and atmosphere light. some researchers focus on the more challenging night-time dehazing problem [93], [94]. In addition to image dehazing, AOD-Net [8], [95] considers the joint interplay effect of dehazing and object detection in an unified framework. The idea is further applied to video dehazing by extending the model into a light-weight video hazing framework [96]. In another recent work [97], the semantic prior is also injected to facilitate video dehazing.

2) Low Light Enhancement: All low-light enhancement methods can be categorized into three ways: hand-crafted methods, Retinex theory-based methods and data-driven methods. Hand-crafted methods explore and apply various image priors to single image low-light enhancement, *e.g.* histogram equalization [98], [99]. Some methods [100], [101] regard the inverted low-light images as hazy images, and enhance the visibility by applying dehazing. The retinex theory-based method [102] is designed to transform the signal components, reflectance and illumination, differently to simultaneously suppress the noise and preserve high-frequency details. Different ways [103], [104] are used to decompose the signal and diverse priors [105]–[108] are applied to realize better light adjustment and noise suppression. Li *et al.* [20] further extended the traditional Retinex model to a robust one with an explicit noise term, and made the first attempt to estimate a noise map out of that model via an alternating direction minimization algorithm. A successive work [21] develops a fast sequential algorithm. Learning based low-light image enhancement methods [22]–[24] have also been studied, where low-light images used for training are synthesized by applying random Gamma transformation on natural normal light images. Some recent works aim to build paired training data from real scenes. In [58], Chen *et al.* introduced a See-in-the-Dark (SID) dataset of short-exposure low-light raw images with corresponding long-exposure reference raw images. Cai *et al.* [109] built a dataset of under/over-contrast and normal-contrast encoded image pairs, in which the reference normal-contrast images are generated by Multi-Exposure image Fusion or High Dynamic Range algorithms. Recently, Jiang *et al.* [25] proposed for the first time an unsupervised generative adversarial network, that can be trained without low/normal-light image pairs, yet generalizing nicely and flexibly on various real-world images.

3) Deraining: Single image deraining is a highly ill-posed problem. To address it, many models and priors are used to perform signal separation and texture classification. These models include sparse coding [110], generalized low rank model [10], nonlocal mean filter [111], discriminative sparse coding [11], Gaussian mixture model [12], rain direction prior [13], transformed low rank model [14]. The presence of deep learning has promoted the development of single image deraining. In [15], [16], deep networks take the image detail layer as their input. Yang *et al.* [17] proposed a deep joint rain detection and removal method to remove heavy rain streaks and accumulation. In [13], a novel density-aware multi-stream densely connected CNN is proposed for joint rain density estimation and removal. Video deraining can additionally make

TABLE I

SUB-CHALLENGE 2.1: IMAGE AND OBJECT STATISTICS OF THE TRAINING/VALIDATION, AND THE HELD-OUT TEST SETS

	#Images	#Bounding Boxes
Training/Validation	4,310	41,113
Test (held-out)	2,987	24,201

TABLE II

SUB-CHALLENGE 2.1: CLASS STATISTICS OF THE TRAINING/ VALIDATION, AND THE HELD-OUT TEST SETS

Categories	*Car*	*Person*	*Bus*	*Bicycle*	*Motorcycle*
RTTS	25,317	11,366	2,590	698	1,232
Test (held-out)	18,074	1,562	536	225	3,804

use of the temporal context and motion information. The early works formulate rain streaks with more flexible and intrinsic characteristics, including rain modeling [10], [112]–[122]. The presence of learning-based method [123]–[129], with improved modeling capacity, brings new progress. The emergence of deep learning-based methods push performance of video deraining to a new level. Chen *et al.* [130] integrated superpixel segmentation alignment, and consistency among these segments and CNN-based detail compensation network into a unified framework. Liu *et al.* [131] presented a recurrent network integrating rain degradation classification, deraining and background reconstruction.

C. Visual Recognition Under Adverse Conditions

A real-world visual detection/recognition system needs to handle a complex mixture of both low-quality and high-quality images. It is commonly observed that, mild degradations, *e.g.* small noises, scaling with small factors, lead to almost no change of recognition performance. However, once the degradation level passes a certain threshold, there will be an unneglected or even very significant effect on system performance. In [132], Torralba *et al.* showed that, there will be a significant performance drop in object and scene recognition when the image resolution is reduced to 32×32 pixels. In [133], the boundary where the face recognition performance is largely degraded is 16×16 pixels. Karahan *et al.* [134] found Gaussian noise with its standard deviation ranging from 10 to 20 will cause a rapid performance decline. In [135], more impacts of contrast, brightness, sharpness, and out-of-focus on face recognition are analyzed.

In the era of deep learning, some methods [3], [136], [137] attempt to first enhance the input image and then forward the output into a classifier. However, this separate consideration of enhancement may not benefit the successive recognition task, because the first stage may incur artifacts which will damage the second stage recognition. In [133], [138], class-specific features are extracted as a prior to be incorporated into the restoration model. In [39], Zhang *et al.* developed a joint image restoration and recognition method based on sparse representation prior, which constrains the identity of the test image and guides better reconstruction and recognition. In [8], Li *et al.* considered dehazing and object detection jointly. These two stage joint optimization methods achieve better performance than previous one-stage methods. In [40], [139], the joint optimization pipeline for low-resolution recognition is examined. In [41], [42], Liu *et al.* discussed the impact of denoising for semantic segmentation and advocated their mutual optimization. Lately, in [140], the algorithmic impact of enhancement algorithms for both visual quality and automatic object recognition is thoroughly examined, on a real image set with highly compound degradations. In our work, we take a further step to consider the joint enhancement and detection in bad weather environments. Three large-scale datasets are collected to inspire new ideas and novel methods in the related fields.

TABLE III

SUB-CHALLENGE 2.2: COMPARISON OF LOW-LIGHT IMAGE UNDERSTANDING DATASETS

Dataset	Training		Testing	
	#Image	#Face	#Image	#Face
ExDark	400	-	209	-
UFDD	-	-	612	-
DarkFace	6,000	43,849	4,000	37,711

III. INTRODUCTION OF UG^{2+} TRACK 2 DATASETS

A. (Semi-)Supervised Object Detection in the Haze

In Sub-challenge 2.1, we use the 4,322 annotated real-world hazy images of the RESIDE RTTS set [71] as the training and/or validation sets (the split is up to the participants). Five categories of objects (car, bus, bicycle, motorcycle, pedestrian) are labeled with tight bounding boxes. We provide another 4,807 unannotated real-world hazy images collected from the same traffic camera sources, for the possible usage of semi-supervised training. The participants can optionally use pre-trained models (*e.g.*, on ImageNet or COCO), or external data. But if any pre-trained model, self-synthesized or self-collected data is used, that must be explicitly mentioned in their submissions, and the participants must ensure all their used data to be public available at the time of challenge submission, for reproduciblity purposes.

There is a held-out test set of 2,987 real-world hazy images, collected from the same sources, with the same classes of objected annotated. Fig. 1 shows the basic statistics of the RTTS set and the held-out set. The held-out test set has a similar distribution of number of bounding boxes per image, bounding box size and relative scale of bounding boxes to input images compared to the RTTS set, but has relatively larger image size. Samples from RTTS set and held-out set can be found in Fig. 2 and Fig. 3.

B. (Semi-)Supervised Face Detection in the Low Light Condition

In Sub-challenge 2.2, we use our self-curated DARK FACE dataset. It is composed of 10,000 images (6,000 for training and validation, and 4,000 for testing) taken in under-exposure condition where human faces are annotated by human with bounding boxes; and 9,000 images taken with the same equipment in the similar environment without human annotations.

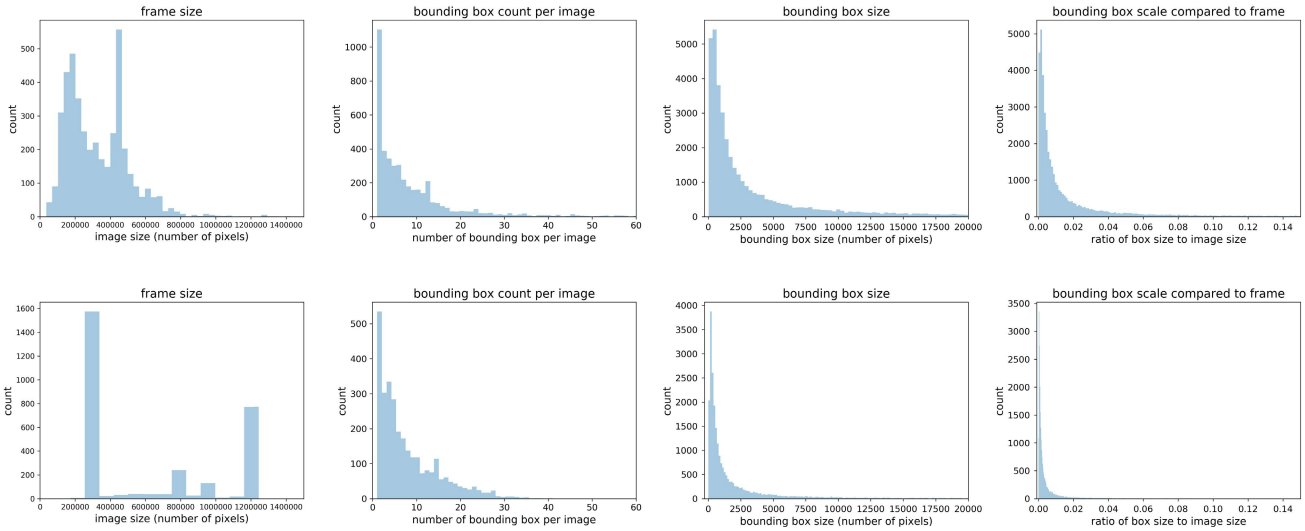


Fig. 1. Sub-challenge 2.1: Basic statistics on the training/validation set (the top row) and the held out test set (the bottom row). The first column shows the image size distribution (number of pixels per image), The second column the bounding box count distribution (number of bounding boxes per image), the third column the bounding box size distribution (number of pixels per bounding box), and the last column the ratios of bounding box size compared to frame size.

Fig. 2. Sub-challenge 2.1: Examples of images in training/validation set (*i.e.*, RESIDE RTTS [71]).

Fig. 3. Sub-challenge 2.1: Examples of images in the held-out test set.

Additionally, we provide a unique set of 789 paired low-light/normal-light images captured in controllable real lighting conditions (but unnecessarily containing faces), which can be optionally used as parts of the training data. The training and evaluation set includes 43,849 annotated faces and the held-out test set includes 37,711 annotated faces. Table III presents a summary of the dataset.

Collection and Annotation: This collection consists of images recorded from Digital Single Lens Reflexes, specifically Sony α6000 and α7 E-mount cameras with different capturing parameters on several busy streets around Beijing, where faces of various scales and poses are captured. The images in this collection are open source content tagged with an Attribution-NonCommercial-NoDerivatives 4.0 International license.[1] The resolution of these images is 1080 $\times$720 (down-sampled from 6K $\times$ 4K). After filtering out those without sufficient information (lacking faces, too dark to see anything, *etc.*), we select 10,000 images for human annotation. The bounding boxes are labeled for all the recognizable faces in our collection. We make the bounding boxes tightly around the forehead, chin, and cheek, using the LabelImg Toolbox.[2] If a face is occluded, we only label the exposed skin region. If most of a face is occluded, we ignore it. For this collection, we observe commonly seen degradations in addition to under-exposure, such as intensive noise. The face number and resolution range distribution are displayed

[1]https://www.jet.org.za/clearinghouse/projects/primted/resources/creative-commons-licence

[2]https://github.com/tzutalin/labelImg

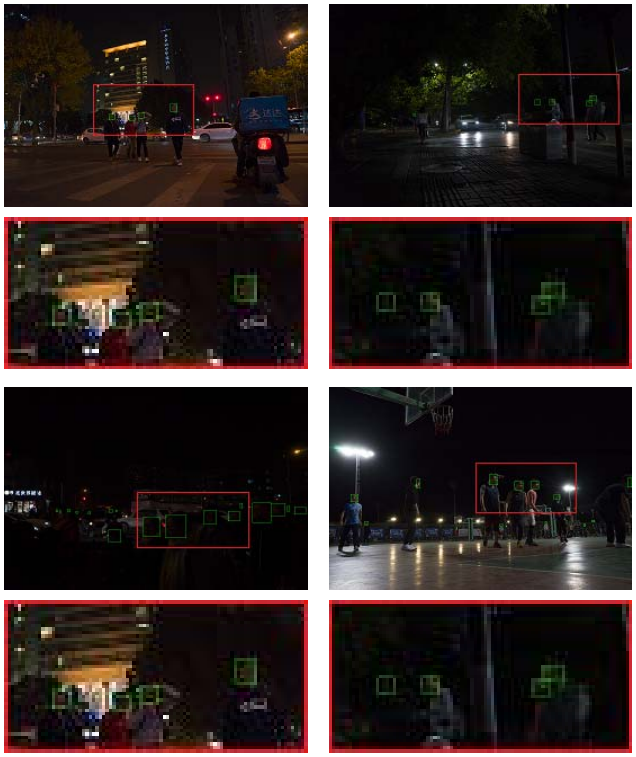


Fig. 4. Sub-challenge 2.2: DARK FACE has a high degree of variability in scale, pose, occlusion, appearance and illumination. The face regions in the red boxes are zoomed-in for better viewing..

in Fig 5. Each annotated image contains 1-34 human faces. The face resolutions in these images range from 1×2 to 335×296. The resolution of most faces in our dataset is below 300 pixel^2 and the the face number mostly falls into the range $[1, 20]$.

C. Zero-Shot Object Detection With Raindrop Occlusions

In Sub-challenge 2.3, we release 1,010 pairs of realistic raindrop images and corresponding clean ground-truths, collected from the real scenes as described in [19], as the training and/or validation sets. Our held-out test set contains 2,495 real rainy images from high-resolution driving videos. As shown in Fig. 6, all images are contaminated by raindrops on camera lens. They are captured in diverse real traffic locations and scenes during multiple drives. We label bounding boxes for selected traffic objects: car, person, bus, bicycle, and motorcycle, which commonly appear on the roads of all images. Most images are of 1920×990 resolution, with a few exceptions of 4023×3024 resolution. The participants are free to use pre-trained models (*e.g.*, ImageNet or COCO) or external data. But if any pre-trained model, self-synthesized or self-collected data is used, that must be explicitly mentioned in their submissions, and the participants must ensure their used data to be public available at the time of challenge submission, for reproducibility purposes.

D. Ranking Criterion

The ranking criteria will be the Mean average precision (mAP) on each held-out test set, with a default Interception-of-Union (IoU) threshold as 0.5. If the ratio of the intersection of a detected region with an annotated object is greater than 0.5, a score of 1 is assigned to the detected region, and 0 otherwise. When mAPs with IoU as 0.5 are equal, the mAPs with higher IoUs (0.6, 0.7, 0.8) will be compared sequentially.

TABLE IV
SUB-CHALLENGE 2.3: OBJECT STATISTICS IN THE HELD-OUT TEST SET

Categories	*Car*	*Person*	*Bus*	*Bicycle*	*Motorcycle*
Test Set	7332	1135	613	268	968

IV. BASELINE RESULTS AND ANALYSIS

For all three sub-challenges, we report results by **cascading off-the-shelf enhancement methods and popular pre-trained detectors**. There has been no joint training performed, hence the baseline numbers are in no way very competitive. We expect to see much performance boosts over the baselines from the competition participants.

A. Sub-Challenge 2.1 Baseline Results

1) Baseline Composition: We test four state-of-the-art object detectors: (1) **Mask R-CNN**[3] [141]; (2) **RetinaNet**[4] [143]; and (3) **YOLO-V3**[5] [144]; (4) Feature Pyramid Network[6] **(FPN)** [145]. We also try three state-of-the-art dehazing approaches: (a) **AOD-Net**[7] [8]; (b) Multi-Scale Convolutional Neural Network (**MSCNN**)[8] [7]; (c) Densely Connected Pyramid Dehazing Network (**DCPDN**)[9] [146]. All dehazing models adopt officially released versions.

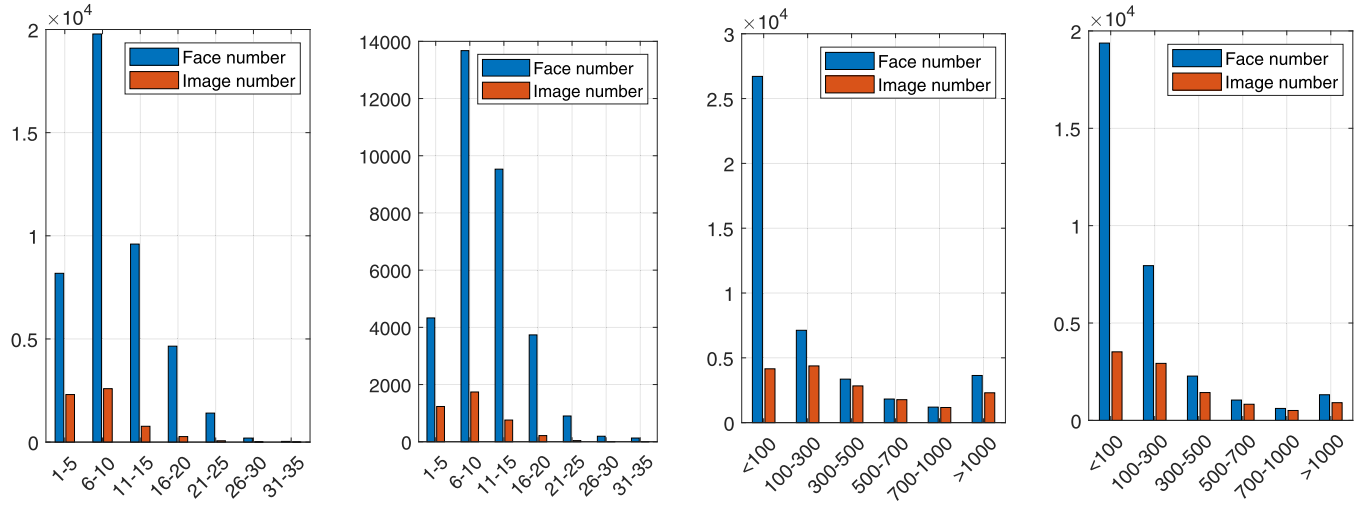
2) Results and Analysis: We evaluate the object detection performance on the original hazy images of RESIDE RTTS set using Mask R-CNN. The detectrons are pretrained on Microsoft COCO, a large-scale object detection, segmentation, and captioning dataset. The detailed detection performance on the five objects can be found in Table V. Results show that without preprocessing or dehazing, the object detectors pretrained on clean images fail to predict a large amount of objects in the hazy image. The overall detection performance has a mAP of only 41.83% using Mask R-CNN and 42.54% using YOLO-V3. Among all the five object categories, person has the highest detection AP, while bus has the lowest AP.

We also compare the validation and test set performance in Table. V. One possible reason for the performance gap between validation and test sets is that the bounding box size of the latter is smaller compared to the former, as showed in Fig. 1 as well as visualized in Fig. 7.

Besides, we analyze the difference between the synthetic haze/rain images and those in real applications. The haze image is generated from the model:

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

[3]https://github.com/matterport/Mask_RCNN
[4]https://github.com/fizyr/keras-retinanet
[5]https://github.com/ayooshkathuria/pytorch-yolo-v3
[6]https://github.com/DetectionTeamUCAS/FPN_Tensorflow
[7]https://github.com/Boyiliee/AOD-Net
[8]https://github.com/rwenqi/Multi-scale-CNN-Dehazing
[9]https://github.com/hezhangsprinter/DCPDN



(a) FN in Train (b) FN in Test (c) FR in Train (d) FR in Test

Fig. 5. Sub-challenge 2.2: Face resolution (FR) and face number (FN) distribution in DARK FACE collections. Image number denotes the number of images belonging to a certain category. Face number denotes the summation number of faces belonging to a certain category.

Fig. 6. Sub-challenge 2.3: Example images from the held-out test set.

where I is the observed hazy image, J is the scene reliance to be recovered. A denotes the global atmospheric light, and t is the transmission matrix assumed correlated with the scene depth. For synthesis hazy images as shown at the top panel of Fig. 8, t is strictly inferred from the results of depth estimation. Therefore, the haze is distributed more homogeneously among different regions and objects, as existing depth estimation techniques are not adaptive enough to accurately estimate the fine-grained depth of each object. Comparatively, the haze layers of real hazy images as shown in the bottom panel of Fig. 8 might be uncorrelated to the scene depth or have more variations for objects with various depths in single hazy images. Besides, the scattering and refraction of light under real hazy condition is different to that in clear environment. The glow effects in front of the vehicles in haze (*e.g.* last examples) makes the vehicle recognition more difficult using pretrained object detection algorithms.

3) Effect of Dehazing: We further evaluate the current state-of-the-art dehazing approaches on hazy dataset, with pre-trained detectors subsequently applied without tuning or adaptation. Fig. 7 shows two examples that dehazing algorithms can improve not only the visual quality of the images but also the detection accuracies. More detection results are included in Table. V. Detection mAPs of dehazed images using DCPDN and MSCNN approaches are 1% higher on average compared to those of hazy images. Eventually, the choice of pre-trained detectors seem to also matter here: Mask R-CNN outperforms the other two detectors on both validation and test sets, before and after dehazing.

Furthermore, as reported in [146] and Table VI, DCPDN has the best SSIM scores while MSCNN has the worst visual quality. However, the detection performance of MSCNN is much better than that of DCPDN.

B. Sub-Challenge 2.2 Baseline Results

1) Baseline Composition: We test four state-of-the-art deep face detectors: (1) Dual Shot Face Detector (**DSFD**) [147][10]; (2) **Pyramidbox** [148][11]; (3) Single Stage Headless Face Detector (**SSH**) [149][12]; (4) **Faster RCNN** [150].[13]

We also include seven state-of-the-art algorithms for light/contrast enhancement: (a) Bio-Inspired Multi-Exposure Fusion (**BIMEF**) [151][14]; (b) **Dehazing** [152][14]; (c) Low-light IMage Enhancement (**LIME**) [107][15]; (d) **MF** [106][14]; (e) Multi-Scale Retinex (**MSR**) [104][14]; (f) Joint Enhancement and Denoising (**JED**) [21][16]; (g) **RetinexNet** [59].[17]

2) Results and Analysis: Fig. 14 (a) depicts the precision-recall curves of the original face detection methods, without enhancement. The baseline methods are trained on WIDER

[10] https://github.com/TencentYoutuResearch/FaceDetection-DSFD
[11] https://github.com/EricZgw/PyramidBox
[12] https://github.com/mahyarnajibi/SSH.git
[13] https://github.com/playerkk/face-py-faster-rcnn
[14] https://github.com/baidut/BIMEF
[15] https://sites.google.com/view/xjguo/lime
[16] https://github.com/tonghelen/JED-Method
[17] https://github.com/weichen582/RetinexNet

Fig. 7. Examples of object detection of hazy images and dehazed images on RESIDE RTTS set. The first column displays the ground truth bounding boxes on hazy images, the second column displays detected bounding box on hazy image using pretrained Mask R-CNN, the right three columns display Mask R-CNN detected bounding boxed on dehazed images using AOD-Net, MSCNN, DCPDN correspondingly.

Fig. 8. Compared with synthetic hazy images in OTS dataset (at the top panel), the haze layers of real hazy images (at the bottom panel) in RTTS dataset might be uncorrelated to the scene depth and have more variations for objects with various depths in single hazy images.

FACE [153],[18] a large dataset with large scale variations in diversified factors and conditions. The results demonstrate that without proper pre-processing or adaptation, the state-of-the-art methods cannot achieve desirable detection rates on DARK FACE. Result examples are illustrated in Fig. 12. The evidences may imply that previous face datasets, though covering variations in poses, appearances, scale, *et al.*, are still insufficient to capture the facial features in the highly under-exposed condition.

We have showed some failure cases of Sub-challenge 2.2. In these results, the faces are falsely detected (false positives and false negatives) due to heavy degradation, small scale, pose variation, occlusion, *etc*. As show in Fig. 9, due to the above mentioned reasons, there are a certain amount of false negative samples caused by heavy degradation, small scale, pose variation, occlusion at the top four panels, respectively, and false positive samples at the bottom panel. DSFD is used as the baseline method. For better visibility, the results shown here are processed by LIME.

3) Effect of Enhancement: We next use the enhancement algorithms to pre-process the annotated dataset and then apply the above two pre-trained face detection methods to the processed data. While the visual quality of the enhanced images is better, as expected, the detectors do perform better. As shown in Fig. 14 (b) and (c), in most instances, the precision of the detectors notably increases compared to that of the data without enhancement. Various existing enhancement methods seem to result in similar improvements here.

[18]http://shuoyang1213.me/WIDERFACE/

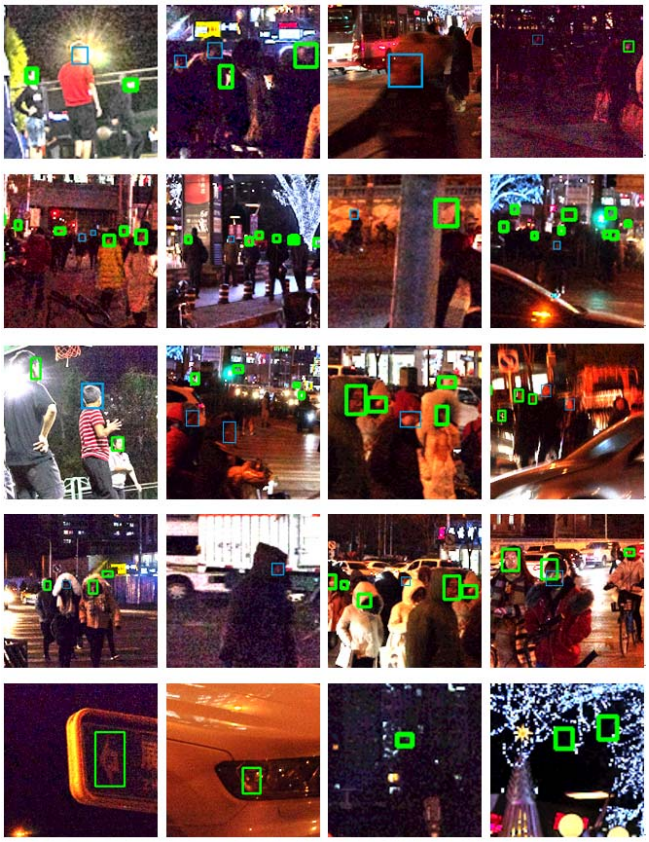


Fig. 9. Failure case analysis even with the model being trained with the proposed training set. Green Box: detection results by the baseline DSFD. Blue boxes: ground truths that are not detected.

Despite being encouraging to see, the overall performance of the detectors still drops a lot compared to normal-light datasets. The simple cascade of low light enhancement and face detectors leave much improvement room open.

C. Sub-Challenge 2.3 Baseline Results

1) Baseline Composition: We use four state-of-the-art object detection models: (1) Faster R-CNN (**FRCNN**) [154]; (2) **YOLO-V3** [144]; (3) **SSD-512** [155]; and (4) **RetinaNet** [143].

TABLE V
DETECTION RESULTS (MAP) ON THE RTTS (TRAIN/VALIDATION DATASET) AND HELD-OUT TEST SETS

mAP			hazy	AOD-Net [8]	DCPDN [7]	MSCNN [148]
validation	∗ RetinaNet [145]	Person	55.85	54.93	56.70	**58.07**
		Car	41.19	37.61	42.68	**42.77**
		Bicycle	39.61	37.80	**36.96**	38.16
		Motorcycle	27.37	23.31	**29.18**	29.01
		Bus	16.88	15.70	16.34	**18.34**
		mAP	36.18	33.87	36.37	**37.27**
	∗ Mask R-CNN [144]	Person	67.52	66.71	67.18	**69.23**
		Car	48.93	47.76	**52.37**	51.93
		Bicycle	**40.81**	39.66	40.40	40.42
		Motorcycle	33.78	26.71	**34.58**	31.38
		Bus	18.11	16.91	18.25	**18.42**
		mAP	41.83	39.55	**42.56**	42.28
	∗ YOLO-V3 [146]	Person	60.81	60.21	60.42	**61.56**
		Car	47.84	47.32	48.17	**49.75**
		Bicycle	41.03	**42.22**	40.18	42.01
		Motorcycle	39.29	37.55	38.17	**41.11**
		Bus	23.71	20.91	**23.35**	23.15
		mAP	42.54	41.64	42.06	**43.52**
	⋄ FPN [147]	Person	51.85	52.35	51.04	**54.50**
		Car	37.48	36.05	37.19	**38.88**
		Bicycle	35.31	35.93	32.57	**37.01**
		Motorcycle	23.65	21.07	22.97	**23.86**
		Bus	12.95	13.68	12.07	**15.83**
		mAP	32.25	31.82	31.17	**34.02**
test	RetinaNet	Person	17.64	18.23	16.65	**19.34**
		Car	31.41	29.30	33.31	**32.97**
		Bicycle	0.42	**0.84**	0.38	0.75
		Motorcycle	1.69	1.37	1.93	**2.03**
		Bus	12.77	13.70	12.07	**15.82**
		mAP	12.79	12.69	12.87	**14.18**
	Mask R-CNN	Person	25.60	26.63	24.59	**27.94**
		Car	39.31	39.71	**42.76**	42.57
		Bicycle	**0.64**	0.52	0.22	0.37
		Motorcycle	**3.37**	2.81	2.83	2.99
		Bus	15.66	15.41	**16.69**	16.55
		mAP	16.92	17.02	17.42	**18.09**
	YOLO-V3	Person	20.64	21.41	21.42	**22.11**
		Car	34.68	33.90	34.52	**35.93**
		Bicycle	0.50	0.38	**0.98**	0.57
		Motorcycle	4.26	4.10	4.72	**5.27**
		Bus	13.55	14.35	13.75	**15.04**
		mAP	14.69	14.83	15.08	**15.78**
	FPN	Person	12.65	12.57	11.13	**14.19**
		Car	30.54	31.24	27.81	**32.68**
		Bicycle	**1.91**	0.39	1.12	0.97
		Motorcycle	**2.25**	1.7	1.96	1.89
		Bus	6.08	7.93	7.39	**8.31**
		mAP	10.69	10.77	9.88	**11.61**

∗ RetinaNet, Mask R-CNN and YOLO-V3 are pretrained on Microsoft COCO dataset.
⋄ FPN using ResNet-101 backbone is pretrained on the PASCAL Visual Object Classes (VOC) dataset.

We employ five state-of-the-art deep learning-based deraining algorithms: (a) JOint Rain DEtection and Removal[19] (**JORDER**) [17]; (b) Deep Detail Network[20] (**DDN**) [15]; (c) Conditional Generative Adversarial Network[21] (**CGAN**) [63]; (d) Density-aware Image De-raining method using a Multistream Dense Network[22] (**DID-MDN**) [13]; and (e) **DeRaindrop**[23] [19]. For fair comparisons, we re-train all deraining algorithms using the same provided training set.

[19]http://www.icst.pku.edu.cn/struct/Projects/joint_rain_removal.html
[20]https://github.com/XMU-smartdsp/Removing_Rain
[21]https://github.com/TrinhQuocNguyen/Edited_Original_IDCGAN
[22]https://github.com/hezhangsprinter/DID-MDN
[23]https://github.com/rui1996/DeRaindrop

TABLE VI
COMPARISON OF HUMAN AND MACHINE VISION QUALITY DIFFERENT METHODS ACHIEVE

Metric	Dataset	Baseline	MSCNN	AOD-Net	DCPDN
SSIM	TestA [140]	-	0.8203	0.8842	0.956
SSIM	TestB [140]	-	0.7724	0.8325	0.8746
MAP	UG2.1-Test	RetinaNet	14.18	12.69	12.87
		Mask R-CNN	18.09	17.02	17.42
		YOLO-V3	15.78	14.83	15.08
		FPN	11.61	10.77	9.88

2) Results and Analysis: Table VII shows mAP results comparisons for different deraining algorithms using different detection models on the held-out test set. Unfortunately, we find that almost all existing deraining algorithms deteriorate the objects detection performance compared to directly

TABLE VII
DETECTION RESULTS (MAP) ON THE HELD-OUT TEST SET

	Rainy	JORDER [17]	DDN [15]	CGAN [63]	DID-MDN [13]	DeRaindrop [64]
FRCNN [156]	16.52	16.97	18.36	**23.42**	16.11	15.58
YOLO-V3 [146]	**27.84**	26.72	26.20	23.75	24.62	24.96
SSD-512 [157]	**17.71**	17.06	16.93	16.71	16.70	16.69
RetinaNet [145]	**23.92**	21.71	21.60	19.28	20.08	19.73

using the rainy images for YOLO-V3, SSD-512, and RetinaNet (The only exception is the detection results by FRCNN). This could be due to those deraining algorithms are not trained towards the end goal of object detection, they are unnecessary to help this goal, and the deraining process itself might have lost discriminative, semantically meaningful true information, and thus hampers the detection performance. In addition, Table VII shows that YOLO-V3 achieves the best detection performance, independent of deraining algorithms applied. We attribute this to the small objects in a relative long distance from the camera in the test set since YOLO-V3 is known to improve small object detection based on multi-scale prediction structure.

V. COMPETITION RESULTS: OVERVIEW & ANALYSIS

The UG^2+ Challenge (Track 2) in conjunction with CVPR 2019 attracted large deals of attention and participation. More than 260 teams registered; among them, 82 teams finished the dry-run and submitted their final results successfully. Eventually, 6 teams were selected as winners (including a winner and a runner-up, for each sub-challenge).

In the following, we review a part of results from those participation teams who volunteer to disclose their technical details. The full leaderboards can be found at the website.[24]

A. Sub-Challenge 2.1: Competition Results and Analysis

A total of seven teams were able to outperform our best baseline numbers (mAP 18.09). The winner and runner-up teams, *HRI_DET* and *superlab403*, achieve record-high mAP results of 52.71 and 49.22, respectively. All teams used deep learning solutions. In addition to using most sophisticated networks, several interesting observations could be concluded: i) while many teams went with the dehazing-detection cascade idea (like our baselines, but usually jointly trained), the top-2 winners used end-to-end trained/adapted detection models on the hazy training set, without an (implicit) dehazing module; ii) the utilization of unlabeled data seems to open up much potential, and we believe it should be paid more attention to in the future; and iii) multi-scale testing and ensembling contribute to many performance gains.

As the winner team, *HRI_DET* used Faster R-CNN, with the ImageNet-pretrained backbone of ResNeXt-101 [156] and Feature Pyramid Network (FPN) [145]. The Faster R-CNN was then tuned on the mixed dataset of MS COCO, PASCAL VOC and KITTI, with the common data augmentations. To further boost the performance, the team delved deep into the provided unlabeled hazy set, and adopted semi-supervised learning by using the unlabeled data to train the feature extractor with a reconstruction loss. The team used a batch size of 8 and trained the network using 8 Tesla M40 GPUs for 30,000 iterations with an initial learning rate of 0.005. They also found it helpful to apply stochastic weight averaging (SWA) [157] to aggregate several checkpoint models in one training pass, and further to ensemble multiple models (by averaging model weights) obtained from different training passes (*e.g.* with different learning rate schedulers). During inference, a three-scale testing is performed by resizing images to 1333×1000, 1000×750 and 2100×600. The third scale is applied to a closer view of the image, by cropping the foreground region defined as the bounding box of all high-confidence predictions with the first scale.

The runner-up team *superlab403* chose the Cascade R-CNN [158] baseline, and also replaced the original ResNet-101 backbone with ResNeXt-101. The team analyzed the distribution of target aspect ratio from the training set, and selected four new anchor ratios (0.8, 1.7, 2.6, 3.7) by k-means. The team also did a (well-appreciated) label cleaning effort. Data augmentations such as blur, illumination change and color perturbations were adopted in training. The model was trained with an SGD optimizer; the initial learning rate was set as 0.0025, then being decayed by a factor of 0.1 at epochs 8, 11, 21 and 41 (total training epoch number 50).

Other teams have each developed their interesting solutions. For example, the *Mt. Star* team (ranked No. 3, mAP 31.24) adopted a sequential cascade of the dehazing model (DehazeNet [6]) and the detection model (Faster-RCNN), each first pre-trained on their own and then jointly tuned end-to-end on the training set. Multi-scale testing was adopted. The *ilab* team (ranked No. 6, mAP 19.15) also referred to the dehazing-detection cascade idea, but using DeblurGAN [159] (re-trained on haze data) for the dehazing model and Yolo-V2 [144] for the detection model, with a content loss.

B. Sub-Challenge 2.2: Competition Results and Analysis

A total of three teams were able to outperform our best baseline numbers (mAP 39.30). The winner and runner-up teams, *CAS-Newcastle* and *CAS_NEU*, achieve high mAP results of 62.45 and 61.84, respectively. Similarly to Sub-challenge 2.1, all teams used deep learning solutions; yet interestingly, the most successful solutions are based on enhancement-detection cascades, showing a different trend with Sub-challenge 2.1.

The *CAS-Newcastle* and *CAS_NEU* team adopted cascades of low-light enhancement (MSRCR [104]) and detection

[24] http://www.ug2challenge.org/leaderboard19_t2.html

(Selective Refinement Network [142] / RetinaNet [143]) models, where the detection models were directly trained on the enhancement models' preprocessed outputs. The *SCUT-CVC* team (ranked No. 7, mAP 35.18) found tone mapping [160] to be an impressively effective pre-processing, on top of which they tuned two DSFD detectors (with VGG-16 and ResNet-152 backbones), whose results were ensembled by late fusion. The *PHI-AI* team (ranked No. 7, mAP 29.95) adopted a U-Net [161] enhancer and a DSFD detector. The *tjfirst* team (ranked No. 12, mAP 26.50) referred to a more sophisticated enhancement module (first enhancing illumination by LIME [107], then super-resolving by DPSR [162], ended by denoising with BM3D [4]), followed by aggregating the DSFD-detection results on the original and enhanced images.

C. Sub-Challenge 2.3: Competition Results and Analysis

Different from the first two sub-challenges, Sub-challenge 2.3 is substantially more difficult due to its "zero-shot" nature. Typical solutions that we see from the challenge teams include deraining + detection cascades; as well as ensembling multiple pre-trained detectors (*e.g.*, the *CAS-Newcastle-TUM* team). Unfortunately but not too surprisingly, none of the participation teams was able to outperform our baseline. That concurs with the conclusion drawn from the recent benchmark work [73]: *"Perhaps surprisingly at the first glance, we find that almost all existing deraining algorithms will deteriorate the detection performance compared to directly using the rainy images..." "No existing deraining method seems to directly help detection. That may encourage the community to develop new robust algorithms to account for high-level vision problems on real-world rainy images. On the other hand, to realize the goal of robust detection in rain does not have to adopt a de-raining preprocessing; there are other domain adaptation type options..."*.

In fact, our Sub-challenge 2.3 is more difficult and challenging. There is no training set that is close to the testing set provided, which fails all submitted methods. Therefore, the problem is closer to zero-shot and unsupervised learning problem. Existing open-source paired rain image training set, *e.g.* Rain800 [63] and raindrop dataset [19], usually consider only one kind of rain degradation, i.e. rain streak or raindrop. Based on a benchmark paper [71], the synthetic and captured paired images rely on three rain models. The generated rain images are not visually authentic and close to the real captured ones. Their background layers of these images are clear and the objects in these images have their appropriate sizes and locations as shown in Fig. 10.

The testing rain images in our Sub-challenge 2.3 are collected in real driving or surveillance scenarios. They come with other degradation such as rain accumulation, blurring, reflectance, occlusion *etc*. as shown in Fig. 11, which causes the domain shift problem and makes the related restoration and detection tasks harder.

Our challenging results further confirm that, the existing open-source paired datasets are poorly related when it comes to task purposes, *e.g.* object detection, in real scenario. Since driving and surveillance are representative of real application scenarios where deraining may be desired, this sub-challenge is well-worth exploration. We are intended to extend this challenge to next year and look for better deraining algorithms to be proposed.

Fig. 10. Samples of open-source paired rain image training set only including the rain streak and raindrop-related degradation.

Fig. 11. Real rain images captured in driving or surveillance cases have degradation of rain accumulation, blurring, reflectance and occlusions.

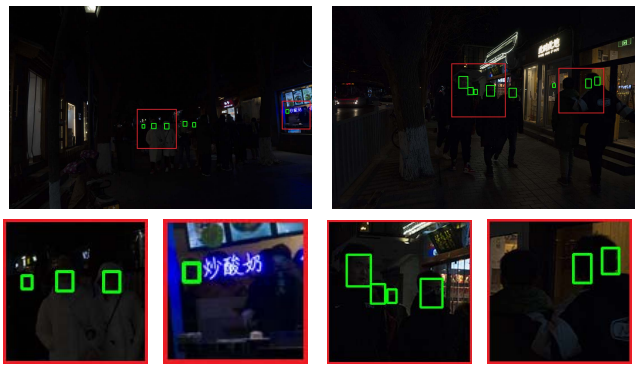
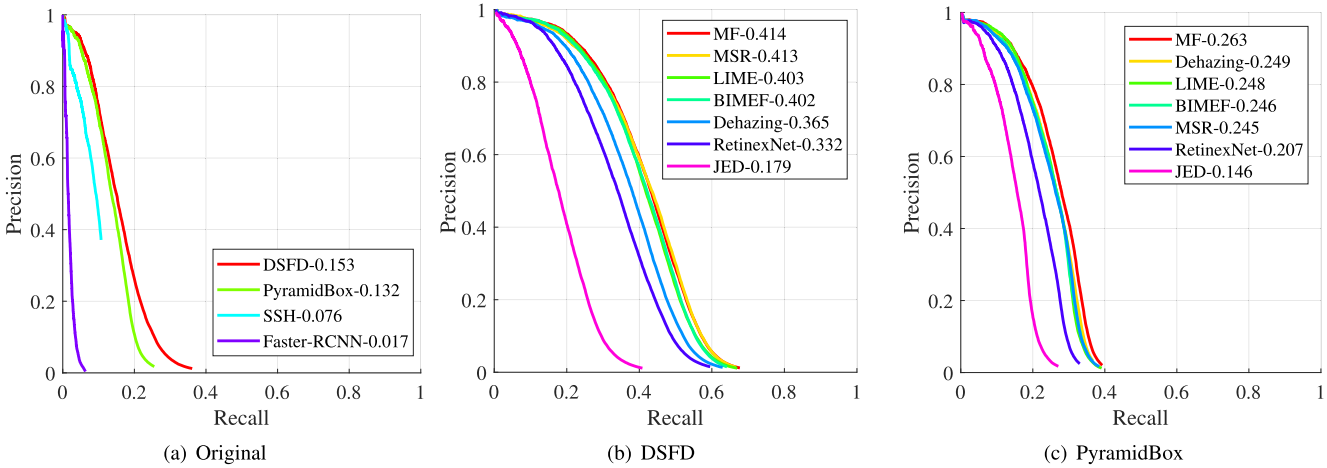


Fig. 12. Sample face detection results of pretrained baseline on the original images of the proposed DARK FACE dataset. The face regions in the red boxes are zoomed-in for better viewing.

Fig. 13. Sample face detection results of pretrained baseline on the enhanced images of the proposed DARK FACE dataset. The face regions in the red boxes are zoomed-in for better viewing.



(a) Original (b) DSFD (c) PyramidBox

Fig. 14. Evaluation results of pretrained baseline on original and enhanced images of the proposed DARK FACE dataset.

VI. Discussions and Future Directions

In this challenge, the submitted solutions and analysis results provide rich experiences, meaningful observations and insights, as well as potential future directions:

- Deep learning methods are preferred by all participants in their submitted solutions. Except that in Sub-challenge 2.2, some teams choose to apply hand-crafted low-light enhancement methods as pre-processing, other submitted methods are deep learning-based as they are flexible to be tuned to improve the performance based on the given task (dataset).
- In Sub-challenge 2.1, the top-2 winners make efforts in exploring the potential of unlabeled data, via semi-supervised learning with a reconstruction loss to guide the reconstruction of the full-picture from the intermediate feature. This strategy leads to performance improvement, which shows that it should be paid more attention to in future.
- For different tasks and techniques used to tackle the problems, the best choices of the framework might be different. In our challenge, the winner of Sub-challenge 2.1 uses a one-step detection scheme (without an implicit dehazing module) while the winner of Sub-challenge 2.2 takes the cascade of enhancement and detection.
- Some teams, *e.g. tjfirst*, report the effectiveness to apply a sequential enhancement process to remove mixed degradation, such as under-exposure, low-resolution blurring, and noise. It also shows a valuable path that is worth further exploration.
- Separate consideration of enhancement and detection might lead to deteriorated performance in detection. In Table V, the offline dahazing operation is first applied by AOD-Net, DCPDN, MSCNN. The results are taken as the input of object detection methods, *i.e.* RetinaNet, Mask R-CNN, YOLO-V3, FPN. Many combination groups of dehazing operation and object detection method obtain inferior results than directly applying object detection without any pre-enhancement, such as (AOD-Net, RetinaNet) for person category ($54.93 < 55.85$) and (AOD-Net, Mask R-CNN), (AOD-Net, Mask R-CNN), (AOD-Net, Mask R-CNN) for all categories ($2.81, 2.83, 2.99 < 3.37$). In Table VII, YOLO-V3, SSD-512 and RetinaNet generate worse object detection results if pre-deraining is applied.
- The existing results show the difficulty of three tasks we set. In Sub-challenge 2.1 and 2.2, the winners only achieve the results below 65 MAP, much inferior to the performance on datasets with degradation. In Sub-challenge 2.3, no participates achieve the results superior to the baseline results. The results show that, our datasets are challenging and there is still large room for further improvement.

VII. Conclusion

As concurred by most teams in the post-challenge feedbacks, it is widely agreed that the three sub-challenges in he UG^{2+} challenge 2019 Track 2 represent a very difficult, under-explored, yet high meaningful class of computer vision problems in practice. While some promising progress has been witnessed from the large volume of team participation, there remains large room to be improved. Through organizing this challenge, we expect to evoke a broader attention from the research community to address these challenges, which are barely covered by previous benchmarks. We look forward to making UG^{2+} a recurring event and also evolving/updating our problems and datasets every year.

Acknowledgment

Wenhan Yang and Ye Yuan helped prepare the dataset proposed for the UG2+ Challenges, and were the main responsible members for UG2+ Challenge 2019 (Track 2) platform setup and technical support. Wenqi Ren, Jiaying Liu, Walter J. Scheirer, and Zhangyang Wang were the main organizers of the challenge and helped prepare the dataset, raise sponsors, set up evaluation environment, and improve the technical submission. Other authors are the group members of winner teams in UG2+ challenge Track 2 contributing to the winning methods.

Wenhan Yang and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: yangwenhan@pku.edu.cn; liujiaying@pku.edu.cn).

Ye Yuan and Zhangyang Wang are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: ye.yuan@tamu.edu; atlaswang@tamu.edu).

Wenqi Ren is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100864, China (e-mail: rwq.renwenqi@gmail.com).

Walter J. Scheirer is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: walter.scheirer@nd.edu).

Taiheng Zhang is with the Department of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: thzhang@zju.edu.cn).

Qiaoyong Zhong, Di Xie, and Shiliang Pu are with the Hikvision Research Institute, Hangzhou 310051, China (e-mail: zhongqiaoyong@hikvision.com; xiedi@hikvision.com; pushiliang.hri@hikvision.com).

Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, and Chen Hong are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361001, China (e-mail: zhengyuqiang@stu.xmu.edu.cn; yyqu@xmu.edu.cn; xyhong@stu.xmu.edu.cn; liang2chen@stu.xmu.edu.cn; 23020181154224@stu.xmu.edu.cn; 786343641@qq.com).

Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, and Pengfei Wan are with the Mtlab, Meitu Inc., Beijing 100080, China (e-mail: jh1@meitu.com; ysy2@meitu.com; ly33@meitu.com; qxc@meitu.com; wpf@meitu.com).

Shuai Zheng, Minhui Zhong, Lingzhi He, Yao Zhao, and Zhenfeng Zhu are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: ericzheng1997@163.com; zylayee@163.com; 19112002@bjtu.edu.cn; yzhao@bjtu.edu.cn; zhfzhu@bjtu.edu.cn).

Taiyi Su is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: tysu@tongji.edu.cn).

Yandong Guo is with Xpeng Motors, Beijing 100080, China (e-mail: guoyd@xiaopeng.com).

Jinxiu Liang, Tianyi Chen, Yuhui Quan, and Yong Xu are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: cssherryliang@mail.scut.edu.cn; csttychen@mail.scut.edu.cn; csyhquan@scut.edu.cn; yxu@scut.edu.cn).

Jingwen Wang is with Tencent AI Lab, Shenzhen 518000, China (e-mail: jwongwang@tencent.com).

Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li and Feng Lu are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: bliu03@buaa.edu.cn; liuxin95@buaa.edu.cn; susiesun@buaa.edu.cn; 16061065@buaa.edu.cn; lixiaochuan822@gmail.com; lufeng@buaa.edu.cn).

Lin Gu is with RIKEN AIP, Tokyo 103-0027, Japan (e-mail: lin.gu@riken.jp).

Shengdi Zhou and Cong Cao are with the School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China.

Shifeng Zhang, Chubing Zhuang, and Zhen Lei are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shifeng.zhang@nlpr.ia.ac.cn; chubin.zhuang@nlpr.ia.ac.cn; zlei@nlpr.ia.ac.cn).

Cheng Chi, Kai Wang, Jiangang Yang, and Likun Qin are with the University of Chinese Academy of Sciences, Beijing 100049, China (chicheng15@mails.ucas.ac.cn; wk2008cumt@163.com; yangjiangang18@mails.ucas.ac.cn; qinlikun19@ucas.ac.cn).

Stan Z. Li is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Engineering, Westlake University, Hangzhou 310024, China (e-mail: szli@nlpr.ia.ac.cn).

Shizheng Wang is with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China, and also with the Chinese Academy of Sciences R&D Center for Internet of Things, Wuxi 214200, China (e-mail: shizheng.wang@foxmail.com).

Ruizhe Liu is with Sunway-AI Co., Ltd., Zhuhai 519000, China, and also with the Chinese Academy of Sciences R&D Center for Internet of Things, Wuxi 214200, China (e-mail: liuruizhe@ciotc.org).

Dong Yi is with Winsense Inc., Beijing 100080, China (e-mail: yidong@winsense.ai).

Zheming Zuo is with the Department of Computer Science, Durham University, Durham 46390, U.K. (e-mail: zheming.zuo@durham.ac.uk).

Jianning Chi, Huan Wang, and Yixiu Liu are with Northeastern University, Shenyang 110819, China (e-mail: chijianning@mail.neu.edu.cn; 1870692@stu.neu.edu.cn; yixiu9713@foxmail.com).

Xingyu Gao is with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China, and also with Beijing Jiaotong University, Beijing 100044, China (e-mail: gaoxingyu@ime.ac.cn).

Zhenyu Chen is with the Big Data Center, State Grid Corporation of China, Beijing 100031, China, and also with the China Electric Power Research Institute, Beijing 100031, China (e-mail: czy9907@gmail.com).

Chang Guo, Yongzhou Li, and Huicai Zhong are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: guochang@ime.ac.cn; liyongzhou@ime.ac.cn; zhonghuicai@ime.ac.cn).

Jing Huang, Heng Guo, and Jianfei Yang are with Nanyang Technological University, Singapore 639798 (e-mail: jhuang027@e.ntu.edu.sg; hguo007@e.ntu.edu.sg; yang0478@e.ntu.edu.sg).

Wenjuan Liao is with The Australian National University, Canberra, ACT 0200, Australia (e-mail: wenjuan.liao@anu.edu.au).

Liguo Zhou and Mingyue Feng are with the Department of Informatics, Technical University of Munich, 85748 Garching, Germany (e-mail: liguo.zhou@tum.de; mingyue.feng@tum.de).

REFERENCES

[1] Y. Zhang, S. Kwong, and S. Wang, "Machine learning based video coding optimizations: A survey," *Inf. Sci.*, vol. 506, pp. 395–423, Jan. 2020.

[2] P. Zhu *et al.*, "VisDrone-DET2018: The vision meets drone object detection in image challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 437–468.

[3] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] 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.

[7] 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. (ECCV)*, 2016, pp. 154–169.

[8] 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. (ICCV)*, Oct. 2017, pp. 4780–4788.

[9] 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.

[10] Y.-L. Chen and C.-T. Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1968–1975.

[11] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3397–3405.

[12] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2736–2744.

[13] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 695–704.

[14] Y. Chang, L. Yan, and S. Zhong, "Transformed low-rank model for line pattern noise removal," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1735–1743.

[15] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1715–1723.

[16] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2944–2956, Jun. 2017.

[17] 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. (CVPR)*, Jul. 2017, pp. 1685–1694.

[18] W. Yang, R. T. Tan, J. Feng, J. Liu, S. Yan, and Z. Guo, "Joint rain detection and removal from a single image with contextualized deep networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 28, 2019, doi: 10.1109/TPAMI.2019.2895793.

[19] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[20] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust Retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[21] X. Ren, M. Li, W.-H. Cheng, and J. Liu, "Joint enhancement and denoising method via sequential decomposition," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[22] J. Yang, X. Jiang, C. Pan, and C.-L. Liu, "Enhancement of low light level images with coupled dictionary learning," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 751–756.

[23] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[24] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net: Low-light image enhancement using deep convolutional network," 2017, *arXiv:1711.02488*. [Online]. Available: http://arxiv.org/abs/1711.02488

[25] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," 2019, *arXiv:1906.06972*. [Online]. Available: http://arxiv.org/abs/1906.06972

[26] Y. Hu, W. Yang, and J. Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 1–8.

[27] J. Liu, S. Xia, and W. Yang, "Deep reference generation with multi-domain hierarchical constraints for inter prediction," *IEEE Trans. Multimedia*, early access, Dec. 23, 2019, doi: 10.1109/TMM.2019.2961504.

[28] J. Liu, S. Xia, W. Yang, M. Li, and D. Liu, "One-for-all: Grouped variation network-based fractional interpolation in video coding," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2140–2151, May 2019.

[29] X. Zhang, W. Yang, Y. Hu, and J. Liu, "DMCNN: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 390–394.

[30] D. Wang, S. Xia, W. Yang, Y. Hu, and J. Liu, "Partition tree guided progressive rethinking network for in-loop filtering of HEVC," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2671–2675.

[31] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "COVERAGE—A novel database for copy-move forgery detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 161–165.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.

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

[35] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[36] D. Bradbury. *How Autonomous Vehicles Will Navigate Bad Weather Remains Foggy*. Forbes. Accessed: Nov. 29, 2016. [Online]. Available: https://www.forbes.com/sites/centurylink/2016/11/29/how-autonomous-vehicles-will-navigate-bad-weather-remains-foggy/#2c9b3d248662

[37] A. Ashe. *Ticketbuster: Snow Causes Red-Light Camera to Issue Ticket in Error*. Wtop. Accessed: Apr. 18, 2014. [Online]. Available: https://wtop.com/news/2014/04/ticketbuster-snow-causes-red-light-camera-to-issue-ticket-in-error/

[38] Y. Liu *et al.*, "Improved techniques for learning to dehaze and beyond: A collective study," 2018, *arXiv:1807.00202*. [Online]. Available: http://arxiv.org/abs/1807.00202

[39] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 770–777.

[40] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," 2017, *arXiv:1712.07732*. [Online]. Available: http://arxiv.org/abs/1712.07732

[41] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," 2018, *arXiv:1809.01826*. [Online]. Available: http://arxiv.org/abs/1809.01826

[42] D. Liu, B. Wen, X. Liu, Z. Wang, and T. S. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," 2017, *arXiv:1706.04284*. [Online]. Available: http://arxiv.org/abs/1706.04284

[43] B. Cheng *et al.*, "Robust emotion recognition from low quality and low bit rate video: A deep learning approach," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 65–70.

[44] L. Stasiak, A. Pacut, and R. Vincente-Garcia, "Face tracking and recognition in low quality video sequences with the use of particle filtering," in *Proc. 43rd Annu. Int. Carnahan Conf. Secur. Technol.*, Oct. 2009, pp. 126–133.

[45] V. Vasek, V. Franc, and M. Urban, "License plate recognition and super-resolution from low-resolution videos by convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 132.

[46] Y.-L. Tian, "Evaluation of face resolution for expression analysis," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 82.

[47] C. Shan, S. Gong, and P. W. McOwan, "Recognizing facial expressions at low resolution," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2005, pp. 330–335.

[48] R. Prabhu, X. Yu, Z. Wang, D. Liu, and A. Jiang, "U-finger: Multi-scale dilated convolutional network for fingerprint image denoising and inpainting," 2018, *arXiv:1807.10993*. [Online]. Available: http://arxiv.org/abs/1807.10993

[49] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 135.

[50] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.*, 2012, pp. 711–730.

[51] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[52] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[53] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[54] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit. Workshop*, Jul. 2017, pp. 1110–1121.

[55] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *Proc. 1st Int. Workshop coMics Anal., Process. Understand. (MANPU)*, 2016, pp. 1–5.

[56] J. Xu, H. Li, Z. Liang, D. Zhang, and L. Zhang, "Real-world noisy image denoising: A new benchmark," 2018, *arXiv:1804.02603*. [Online]. Available: http://arxiv.org/abs/1804.02603

[57] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1586–1595.

[58] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3291–3300.

[59] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 155.

[60] Y. Zhang, L. Ding, and G. Sharma, "HazeRD: An outdoor scene dataset and benchmark for single image dehazing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3205–3209.

[61] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," 2018, *arXiv:1804.05101*. [Online]. Available: http://arxiv.org/abs/1804.05101

[62] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images," 2018, *arXiv:1804.05091*. [Online]. Available: http://arxiv.org/abs/1804.05091

[63] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," 2017, *arXiv:1701.05957*. [Online]. Available: http://arxiv.org/abs/1701.05957

[64] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.

[65] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2227–2234.

[66] X. Zhu, C. C. Loy, and S. Gong, "Video synopsis by heterogeneous multi-source correlation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 81–88.

[67] S. Oh *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.*, Jun. 2011, pp. 3153–3160.

[68] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.

[69] B. Z. Yao, X. Yang, and S.-C. Zhu, "Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks," in *Proc. Energy Minimization Methods Comput. Vis. Pattern Recognit.*, 2007, pp. 169–183.

[70] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: A challenge dataset and baseline results," 2018, *arXiv:1804.10275*. [Online]. Available: http://arxiv.org/abs/1804.10275

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

[72] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Understand.*, vol. 178, pp. 30–42, Jan. 2019.

[73] S. Li *et al.*, "Single image deraining: A comprehensive benchmark analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3833–3842.

[74] Z. Wang, H. Li, Q. Ling, and W. Li, "Robust temporal-spatial decomposition and its applications in video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 3, pp. 387–400, Mar. 2013.

[75] H. Li, Z. Lu, Z. Wang, Q. Ling, and W. Li, "Detection of blotch and scratch in video based on video decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1887–1900, Nov. 2013.

[76] J. Ren, J. Liu, and Z. Guo, "Context-aware sparse decomposition for image denoising and super-resolution," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1456–1469, Apr. 2013.

[77] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang, "D3: Deep dual-domain based fast restoration of JPEG-compressed images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2764–2772.

[78] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3252–3265, Dec. 2018.

[79] 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. (CVPR)*, Jul. 2017, pp. 4003–4011.

[80] 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. (ICCV)*, Oct. 2017, pp. 1077–1085.

[81] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4359–4371, Nov. 2015.

[82] Z. Wang *et al.*, "Self-tuned deep super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–8.

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

[84] D. Liu *et al.*, "Learning temporal dynamics for video super-resolution: A deep learning approach," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3432–3445, Jul. 2018.

[85] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, "Multi-level video frame interpolation: Exploiting the interaction among different levels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1235–1248, Jul. 2013.

[86] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, Aug. 2008.

[87] 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.

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

[89] J. Zhou and F. Zhou, "Single image dehazing motivated by Retinex theory," in *Proc. 2nd Int. Symp. Instrum. Meas., Sensor Netw. Autom. (IMSNA)*, Dec. 2013, pp. 243–247.

[90] D. Nair, P. A. Kumar, and P. Sankaran, "An effective surround filter for image dehazing," in *Proc. Int. Conf. Interdiscipl. Adv. Appl. Comput. (ICONIAAC)*, 2014, p. 20.

[91] L. Kratz and K. Nishino, "Factorizing scene albedo and depth from a single foggy image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1701–1708.

[92] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 263–278, Jul. 2012.

[93] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 226–234.

[94] J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. W. Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7016–7024.

[95] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "An all-in-one network for dehazing and beyond," 2017, *arXiv:1707.06543*. [Online]. Available: http://arxiv.org/abs/1707.06543

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

[97] W. Ren *et al.*, "Deep video dehazing with semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1895–1908, Apr. 2019.

[98] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Müller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Visualizat. Biomed. Comput.*, 1990, pp. 337–345.

[99] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, May 2007.

[100] L. Li, R. Wang, W. Wang, and W. Gao, "A low-light image enhancement method for both denoising and contrast enlarging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3730–3734.

[101] X. Zhang, P. Shen, L. Luo, L. Zhang, and J. Song, "Enhancement and noise reduction of very low light level images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 2034–2037.

[102] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–128, Dec. 1977.

[103] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround Retinex," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 451–462, Mar. 1997.

[104] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[105] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[106] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, Dec. 2016.

[107] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[108] X. Fu, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2782–2790.

[109] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[110] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1742–1755, Apr. 2012.

[111] J.-H. Kim, C. Lee, J.-Y. Sim, and C.-S. Kim, "Single-image deraining using an adaptive nonlocal means filter," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 914–917.

[112] K. Garg and S. K. Nayar, "Photorealistic rendering of rain streaks," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 996–1002, Jul. 2006.

[113] K. Garg and S. K. Nayar, "Detection and removal of rain from videos," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Apr./Jun. 2004, p. 1.

[114] K. Garg and S. K. Nayar, "Vision and rain," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 3–27, Jul. 2007.

[115] K. Garg and S. K. Nayar, "When does a camera see rain?" in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2005, pp. 1067–1074.

[116] X. Zhang, H. Li, Y. Qi, W. Leow, and T. Ng, "Rain removal in video by combining temporal and chromatic properties," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 461–464.

[117] P. Liu, J. Xu, J. Liu, and X. Tang, "Pixel based temporal analysis using chromatic property for removing rain from videos," *Comput. Inf. Sci.*, vol. 2, no. 1, pp. 53–60, 2009.

[118] P. C. Barnum, S. Narasimhan, and T. Kanade, "Analysis of rain and snow in frequency space," *Int. J. Comput. Vis.*, vol. 86, nos. 2–3, pp. 256–274, Jan. 2010.

[119] V. Santhaseelan and V. K. Asari, "Utilizing local phase information to remove rain from video," *Int. J. Comput. Vis.*, vol. 112, no. 1, pp. 71–89, Mar. 2015.

[120] N. Brewer and N. Liu, "Using the shape characteristics of rain to identify and remove rain from video," in *Proc. Joint IAPR Int. Workshops (SPR SSPR)*, 2008, pp. 451–458.

[121] J. Bossu, N. Hautière, and J.-P. Tarel, "Rain or snow detection in image sequences through use of a histogram of orientation of streaks," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 348–367, Jul. 2011.

[122] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, L.-J. Deng, and Y. Wang, "A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2818–2827.

[123] J. Chen and L.-P. Chau, "A rain pixel recovery algorithm for videos with highly dynamic scenes," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1097–1104, Mar. 2014.

[124] A. Tripathi and S. Mukhopadhyay, "A probabilistic approach for detection and removal of rain from videos," *IETE J. Res.*, vol. 57, no. 1, pp. 82–91, 2011.

[125] A. K. Tripathi and S. Mukhopadhyay, "Video post processing: Low-latency spatiotemporal approach for detection and removal of rain," *IET Image Process.*, vol. 6, no. 2, pp. 181–196, Mar. 2012.

[126] W. Ren, J. Tian, Z. Han, A. Chan, and Y. Tang, "Video desnowing and deraining based on matrix decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2838–2847.

[127] M. Li *et al.*, "Video rain streak removal by multiscale convolutional sparse coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6644–6653.

[128] W. Wei, L. Yi, Q. Xie, Q. Zhao, D. Meng, and Z. Xu, "Should we encode rain streaks in video as deterministic or stochastic?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2535–2544.

[129] J.-H. Kim, J.-Y. Sim, and C.-S. Kim, "Video deraining and desnowing using temporal correlation and low-rank matrix completion," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2658–2670, Sep. 2015.

[130] J. Chen, C.-H. Tan, J. Hou, L.-P. Chau, and H. Li, "Robust video content alignment and compensation for rain removal in a CNN framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6286–6295.

[131] J. Liu, W. Yang, S. Yang, and Z. Guo, "Erase or fill? Deep joint recurrent rain removal and reconstruction in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3233–3242.

[132] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[133] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[134] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.

[135] A. Dutta, R. Veldhuis, and L. Spreeuwers, "The impact of image quality on the performance of face recognition," in *Proc. Symp. Inf. Theory Benelux Joint WIC/IEEE Symp. Inf. Signal Process. Benelux*, 2012, pp. 141–148.

[136] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, Jul. 2006.

[137] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 576–584.

[138] P. H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[139] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4792–4800.

[140] R. G. VidalMata *et al.*, "Bridging the gap between computational photography and visual recognition," 2019, *arXiv:1901.09482*. [Online]. Available: http://arxiv.org/abs/1901.09482

[141] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 2980–2988.

[142] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 8231–8238.

[143] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[144] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[145] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[146] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.

[147] J. Li *et al.*, "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5055–5064.

[148] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 812–828.

[149] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4875–4884.

[150] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657.

[151] Z. Ying, G. Li, and W. Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," 2017, *arXiv:1711.00591*. [Online]. Available: http://arxiv.org/abs/1711.00591

[152] X. Dong *et al.*, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.

[153] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[154] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[155] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[156] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[157] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," 2018, *arXiv:1803.05407*. [Online]. Available: http://arxiv.org/abs/1803.05407

[158] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[159] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[160] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 419–426, Sep. 2003.

[161] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[162] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[163] J. Wu, X. Yu, D. Liu, M. Chandraker, and Z. Wang, "DAVID: Dual-attentional video deblurring," in *Proc. IEEE WACV*, 2020.

[164] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurganv2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8878–8887.