Journal of Imaging Science and Technology® 65(4): 040404-1–040404-11, 2021.

HDR4CV: High Dynamic Range Dataset with Adversarial Illumination for Testing Computer Vision Methods

Param Hanji and Muhammad Z. Alam
University of Cambridge, Cambridge, UK

Nicola Giuliani and Hu Chen
Huawei Technologies, Munich, Germany

Rafał K. Mantiuk
University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK
E-mail: rafal.mantiuk@cl.cam.ac.uk

Abstract. *Benchmark datasets used for testing computer vision (CV) methods often contain little variation in illumination. The methods that perform well on these datasets have been observed to fail under challenging illumination conditions encountered in the real world, in particular, when the dynamic range of a scene is high. The authors present a new dataset for evaluating CV methods in challenging illumination conditions such as low light, high dynamic range, and glare. The main feature of the dataset is that each scene has been captured in all the adversarial illuminations. Moreover, each scene includes an additional reference condition with uniform illumination, which can be used to automatically generate labels for the tested CV methods. We demonstrate the usefulness of the dataset in a preliminary study by evaluating the performance of popular face detection, optical flow, and object detection methods under adversarial illumination conditions. We further assess whether the performance of these applications can be improved if a different transfer function is used.* © *2021 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2021.65.4.040404]

1. INTRODUCTION

Computer vision (CV) methods are often trained and evaluated on datasets that contain images obtained in relatively "easy" conditions, in which the illumination is mostly uniform across the scene, and there is little camera noise in the images. The performance of such methods can drop substantially when used with images captured in more realistic conditions, where the illumination can vary substantially across the scene. Commonly encountered problems include false edges produced by shadows, contrast reduction due to glare, and camera noise in the darker parts of the scene. These problems have been recognized and addressed by collecting large datasets with varying illumination conditions [1–3] or by simulating different illumination conditions with computer graphics [4] methods.

In this work, we capture a dataset using controlled camera and lighting setups to evaluate the robustness of CV methods under adversarial illumination conditions. Our dataset is composed of video sequences captured for the same scene but under several different illuminations. The frames were captured from the same camera position with the same scene arrangement, while a set of artificial lights were configured to mimic one of four illumination conditions: an "easy" uniform illumination, low-light night condition, high dynamic range (HDR) condition, and a condition with a bright light source that induces strong glare. The main advantage of this approach is that we can use the "easy" uniform condition to produce labels for any CV method. Subsequently, these labels enable us to measure the relative degradation in performance of the method under other illumination conditions. This saves us the manual work of labeling the dataset for each CV application. The dataset is publicly available at https://doi.org/10.17863/CAM.71285.

We captured video sequences with a CV camera mounted on a motorized camera slider, which let us introduce motion parallax and therefore widen the range of applications that can be addressed. Thus our sequences can be used to evaluate optical flow [5, 6] and global motion compensation [7] methods under challenging illuminations. We provide both linear 16-bit demosaiced RGB images and merged HDR images in the OpenEXR format. The former is representative of a typical CV camera, and the latter can be used to simulate a range of cameras and capture scenarios.

In Section 2, we discuss existing datasets and categorize them according to the target applications. Then, in Section 3, we describe our camera and illumination setup, and in Section 4, the construction of our indoor scenes. Next, in Section 5, we provide a summary of the dataset and describe how each frame is processed. To demonstrate the utility of our dataset, we show how several face detection, object detection, and optical flow methods are affected by adversarial illumination in Section 6.1. Finally, in Section 6.2, we analyze how the choice of the color transfer function (TF) can improve the performance of CV methods in adversarial illumination conditions.

2. RELATED WORK

Publicly available image datasets serve as the main means of evaluating and comparing CV methods. Some established

Received Apr. 29, 2021; accepted for publication July 8, 2021; published online Aug. 6, 2021. Associate Editor: Sophie Triantaphillidou.
1062-3701/2021/65(4)/040404/11/$25.00

Figure 1. The *Street* scene along with the capture and setup. Controllable lights are present on either side of the scene to simulate various lighting conditions. The right image shows the spotlight behind the scene, which shines through the diffuser and serves as a source of glare.

datasets, such as Middlebury stereo [8, 9] and optical flow datasets [10] contain well-illuminated images captured in controlled laboratory conditions. A similar trend can be observed for higher-level detection and recognition tasks, for which commonly used evaluation datasets are COCO [11] for objects and LFW [12], CelebA [13], and FFHQ [14] for faces. Although such datasets are highly influential and essential for evaluating CV methods, they attract criticism since they may not sufficiently reflect the varied illuminations found in real-world scenes.

In-the-wild datasets Very large datasets with a good variation of illumination conditions have been captured with monitoring cameras [2, 3] or car-cabin cameras [15]. However, such datasets are intended for a single application, and extending them to other applications requires tedious manual labeling (e.g., 11 man-months of work in some cases [2]). Our dataset cannot match the size of these specialized datasets but can be used across diverse CV applications.

Multi-illumination datasets Scenes with the same composition but varying illumination can be captured using a motorized photographic flashlight on a camera by taking multiple images [1, 16]. The light bouncing off the walls and the ceiling illuminates objects from different directions, providing a large variation in illumination required for image relighting methods. Such an approach, however, is suitable only for indoor scenes without motion, results in a rather artificial structure of the reflected flashlight, and is unlikely to produce high-contrast HDR illumination. Instead, we carefully compose scenes with multiple light sources to achieve challenging illumination conditions.

Abdelhamed et al. [17] captured the smartphone image denoising dataset (SIDD), a dataset for testing denoising methods intended for smartphone cameras. They captured a series of images for ten different static scenes, each under four conditions by varying the camera gain settings, color temperature, and brightness of the light sources. In contrast, our dataset consists of video sequences with motion parallax, captured with a CV camera, in which the illumination was specifically designed to challenge a range of CV methods.

Tracking and detection under challenging illumination conditions Underexposed and saturated regions pose problems to object detection [18] and tracking methods [19]. To address this, Atoum [18] proposes illumination-aware CNNs (convolution neural networks) to improve object detection. Alismail et al. adapt illumination-invariant binary descriptors to achieve photometric invariance in tracking [19]. Our work provides an evaluation dataset for these and other related problems.

Table I. Configuration of lights used for different illumination conditions. The first two lights were controlled using the DMX512 protocol, allowing a value between 0 and 255. The remaining lights could only be turned on or off using relays connected to an Arduino board.

Light	Night	HDR	Glare	Uniform
DMX area	0	63	255	255
DMX spotlight	10	0	255	0
Photographic light	off	on	off	off
Tunnel/street LEDs	on	off	off	on
Background	on	on	on	on

Relighting Image-based relighting can be used to generate novel images of a scene under arbitrary illumination conditions [20, 21]. However, most relighting methods either require several images of the same scene captured under different lighting, or they are restricted to a single application, such as portrait relighting [22]. Since several images need to be captured anyway, we are better off capturing the images under the desired illumination and avoiding potential artifacts of the relighting method used.

Rendering Large datasets with automatically generated labels can be produced with computer graphics methods [4, 23, 24]. Game engines or offline rendering can be used to render photorealistic scenes in arbitrary illumination conditions. However, obtaining highly realistic rendering results, comparable to camera images, requires a substantial amount of effort by skilled professionals. Furthermore, rendered images tend to differ substantially from those captured by cameras. This may introduce a bias in training, validation and result in a model that underperforms *in the wild* [25]. This problem is addressed by domain adaptation methods [25], which involve training on a mixture of real and computer-generated images in a fully or semi-supervised manner or adversarial training on the source and target domains [26].

3. CAPTURE SETUP

In this section, we discuss the common capture and illumination setup shared across all scenes under different illumination conditions. Figure 1 depicts the various components of the capture setup.

3.1 *Lights*

Each illumination condition was simulated by toggling or dimming a combination of lights according to the configurations detailed in Table I. The lights included a compact spotlight (Cameo Q-SPOT 40 TW) and a

Figure 2. The scenes (rows) captured under different illumination conditions (columns). All images have been gamma encoded ($\gamma = 2.2$) for visualization.

photographic box light (Astora SF 120), both controlled over the DMX512 protocol using the QLC+ software [27]. *Selenium webdriver* [28] was used to control QLC+ software using its web interface. We also used a photographic light with a single LED bulb (Omnilux 18W 1800–3000 K), and smaller LED lights controlled by a custom Arduino board with several relays. Additionally, background illumination was provided by an LED video light (Neewer NL480). We tested all the lights with a custom high-frequency light meter to ensure that they were flicker free up to 8.9 kHz.

Some example images from the captured scenes under each illumination condition are shown in the columns of Figure 2. These conditions are described below:

3.1.1 *Uniform*

For each scene, a uniform condition was obtained by manipulating the lights to illuminate as much of the scene as possible. The camera parameters were selected to produce well-exposed frames with low levels of noise. The CV methods tested in our experiments (Sections 6.1 and 6.2) perform very well on images captured under the uniform condition. Thus, it serves as a reference condition for generating labels used to test other, more challenging illumination conditions. The advantage of such a framework is that we prevent tedious, manual labeling of the data for different applications. We note, however, that the labels obtained are not ground truth labels and their only purpose is to test the relative change in performance in under adversarial illuminations.

3.1.2 *Night*

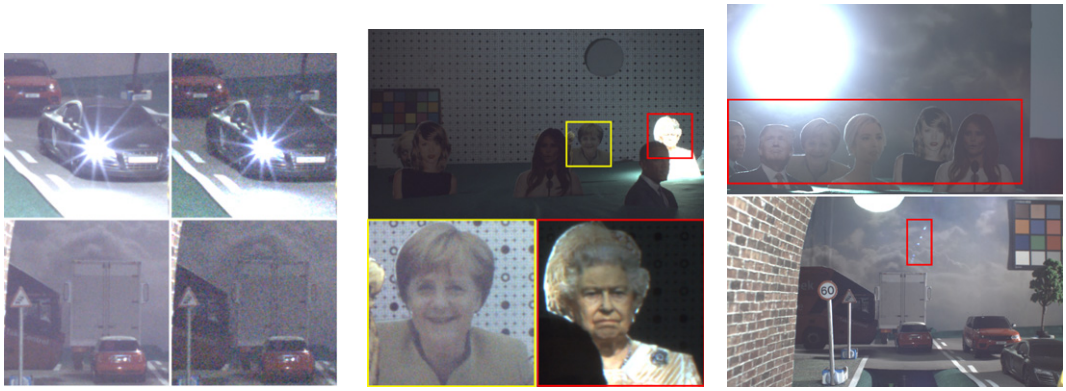
Night illumination simulates low-light conditions by using only a few selected lights and dimming them as needed. In low-light imaging, it is customary to use a high camera gain to capture video, and thus, we used a gain of 16 (the camera's maximum gain). The high gain introduces significant noise in the images as depicted by the zoomed-in patches in Figure 3(a). Using longer exposure times is often not an option for videos, as it results in motion blur and a low frame rate. Another characteristic of the night condition is bright lights such as tunnel lights in the *Tunnel* and street lights in the *Street* scenes.

3.1.3 *High dynamic range*

Here, the lights were arranged to achieve a very high contrast between the dark and bright parts of the scene to simulate HDR illumination conditions (see Fig. 3(b)). Such high contrast was achieved either by illuminating one of the faces with the spotlight (for the *Faces* scene) or by using a foam board to produce sharp shadows from a bright source of light (see Fig. 1) (for other scenes). In the *Tunnel* scene, the ceiling lights of the tunnel were turned off, while photographic lights simulated the bright outdoors. As depicted in Fig. 3(b) and Fig. 2 column (b), dark regions of the scene were barely visible when the image was display encoded using a regular *gamma* encoding. These image regions were also affected by noise much more than their brighter counterparts.

3.1.4 *Glare*

It is often unavoidable to have very strong light sources, such as car headlights or the Sun, in an image. Such strong light sources introduce visible glare, also known as blooming, caused by unwanted scattering and reflections of light inside the camera lens. The glare causes a reduction in contrast in an otherwise well-exposed image, as shown in Fig. 3(c) (top). Depending on the configuration of the scene, we used either



(a) Patches of the night illumination (right column) captured with exposure time $t = 1/150$ secs and gain $g = 16$, compared to uniform illumination (left column).

(b) HDR image along with appropriately exposed insets of dark and bright regions. Notice the noise in the left inset and saturation in parts of the right inset.

(c) Glare or blooming (top) causing contrast reduction and lens flare (bottom). Notice that flare may occur due to lights outside the field of view.

Figure 3. Gamma encoded ($\gamma = 2.2$) images of the dataset showing the various problems or imperfections in images under non-ideal but realistic illumination conditions.

three bright LED bulbs fixed in a photographic frame and connected to a relay, or the compact spotlight controlled by DMX512 protocol to induce glare in the images (see Fig. 1 right).

Lens flare is another problem caused by stray reflections. Lens flare can be seen in Fig. 3(c) (bottom) and is caused by the tunnel lights present in the scene.

3.2 *Camera and Lens*

We captured the scenes using an IDS UI-3860CP-C-HQ computer vision camera, which has a Sony IMX290 1/2.8" CMOS sensor of resolution 1936 × 1096 pixels and pixel size 2.9 µm. The camera was remotely controlled to capture 12-bit RAW images. For each video frame, we captured a stack of 13 images with increasing exposure times with a distance of 1 stop between them. A higher gain of 16 was used for some conditions such as night and HDR. The final images selected for each illumination condition depend on the scene-specific lighting configuration and lens aperture used. These are selected from the captured exposure stacks. An advantage of the captured exposure stacks is that the HDR scenes can be accurately reconstructed. Subsequently, pretrained generative models [29] or calibrated camera parameters [30, 31] can be used to simulate other cameras, generating additional realistic images of the same scenes captured with different camera settings.

Our scenes were captured using one of three lenses depending on the specific camera motion and illumination condition. The different lenses used were:

- Narrow: Fujifilm HF25HA-1B with focal length 25 mm and the effective field of view (accounting for the sensor crop) of 14.6° × 9.78°
- Medium: Navitar HR973NCN with focal length 8 mm and the effective field of view of 43.7° × 29.9°
- Wide: Wide-angle Navitar MVL4WA with focal length 3.5 mm and the effective field of view of 85° × 62.9°

All lenses have a maximum aperture of f/1.4. However, in most captures, we set a much smaller aperture to ensure a sufficiently large depth of field. The exact lens used varied for each sequence and is discussed in Section 4.

3.3 *Motorized Camera Slider*

The camera mounted on a camera slider was powered by a stepper motor driver (Wantai DQ542MA) and controlled with a custom Arduino board connected to a PC. To ensure that the frames were captured from the same viewpoint for each illumination condition, we cycled through all illumination conditions before moving the camera to the next position on the slider. We captured each scene from 100 different camera positions, simulating smooth camera motion. The total length of the movement was 808.89 mm and thus, the distance between each frame of the sequence was 8.09 mm. We positioned the slider parallel, perpendicular, or diagonal to the background plane of the physical scene, depending on the scenario.

4. SCENES

Since we required full control of the lights and needed to capture the scenes over several hours, we had to simulate semi-realistic scenarios in the lab. The scenes consisted of printed foam board cutouts and models in 1:24 scale. All objects were placed at different distances from the camera to introduce parallax. The background was made out of a large foam board with printed photographs or patterns glued on it. All the scenes also included X-Rite Classic Color Checker Chart used for white balance and color calibration.

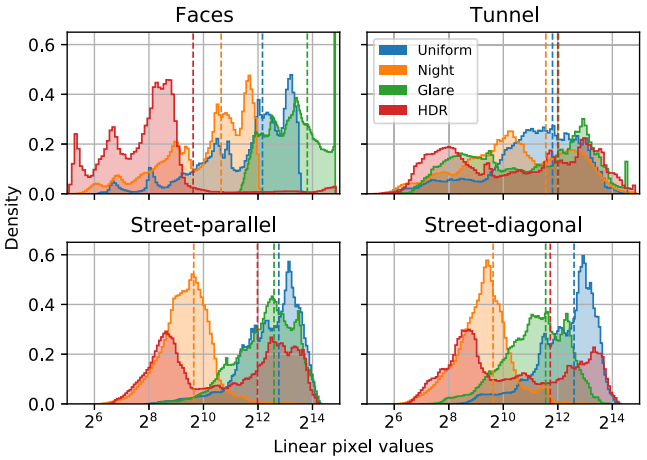


Figure 4. Distribution of pixel values under different illumination conditions for all scenes in the dataset. Dashed vertical lines show the average pixel values under each illumination for each scene.

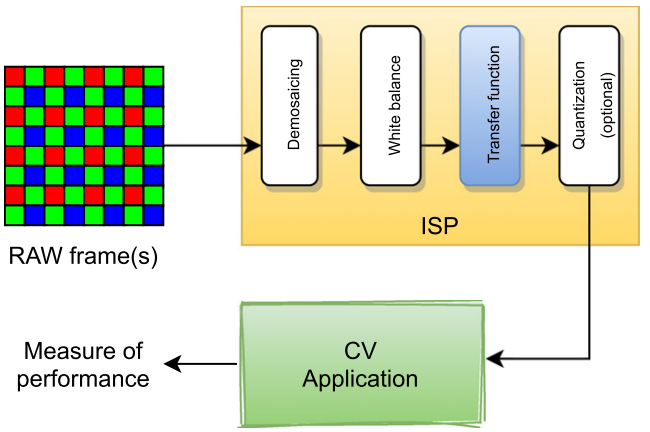


Figure 5. The ISP pipeline, used for evaluating the performance of CV methods. The intermediate frames are processed by a TF, followed by an optional quantization. This serves as an input to the CV application being tested.

4.1 *Faces*

Face detection and recognition are important and well-studied problems. Most face datasets [12–14] contain prominent, well-exposed faces under ideal illumination conditions. To evaluate the robustness of state-of-the-art face detection and recognition methods, we constructed a scene composed of cardboard cutouts of the faces of popular figures. We also included cutouts of a tiger and a gorilla to introduce the possibility of false positives. The source of glare was introduced by cutting a circular hole in the foam board that supported the background and covering it with a diffuser film. As depicted in Fig. 1, the compact spotlight was placed on the other side of the foam board and directed toward the camera. For this scene, the camera slider moved horizontally, parallel to the scene. The dynamic range of the scene was increased by pointing the focused spotlight on one of the cutouts (the Queen in Fig. 3(b)) while keeping the background light dim. The lens with medium focal length was used for HDR and night conditions as it allowed the camera to be placed closer to the scene. However, this lens produced images with significant lens flare in the presence of a bright source of light. For this reason, the narrow-angle lens was used for the glare condition and the camera was placed at a larger distance from the scene. Since we had to use a different lens and also move the spotlight, we captured a separate uniform illumination condition, which served as a reference for the glare condition.

4.2 *Tunnel*

This scene simulated a camera mounted inside a car on a busy road. The motion of the camera slider simulated the movement of the car as it exited a tunnel and approached an intersection. Unlike the other scenes, the camera moved in the direction it was facing, perpendicular to the plane of the background. We used a wide-angle lens, similar to those used in dash cameras. The scene consisted of objects built from cardboard and foam board cutouts (the tunnel, truck, bus, traffic signs, etc.), models of cars in 1:24 scale, and a model of a tree. We modified one of the cars so that it had its headlights on during the capture. For the night illumination condition, the tunnel was illuminated with controllable LED lights.

This scene had a large dynamic range as it included low-luminance regions inside the tunnel and bright regions outside. Lens flare caused by the headlights of oncoming cars and tunnel lights provided additional adversarial illumination.

4.3 *Street*

The final scene depicted a crowded street with cars, a bus, and pedestrians. For the night condition, the street was partially illuminated with small model lamp posts, controlled by the Arduino board. The pedestrians were model people in 1:25 scale from a model train collection (the closest matching scale). The camera moved either along the street (labeled *Street-parallel*) or on a diagonal 45° path (labeled *Street-diagonal*), simultaneously moving along and toward the street. The source of glare had the shape of a set of construction-site lights with the DMX spotlight behind the diffuser film.

5. PROCESSING OF THE FRAMES

We prepared linear color camera frames, stored as 16-bit PNG files, and HDR frames stored as OpenEXR images. The RAW camera frames were first demosaiced using DDFAPD [32] and then white balanced. To perform white balance, we multiplied RGB values so that all color channels had equal intensity for the white patch in the color checker chart.

The 16-bit PNG images were created by selecting an appropriate exposure from the captured exposure stacks. For every scene and illumination, we selected the exposure that avoided saturated pixels, and hence, was representative of typical capture conditions. For example, we chose sequences with high gain for night conditions, and exposures with high contrast and almost no saturation for the HDR scene.

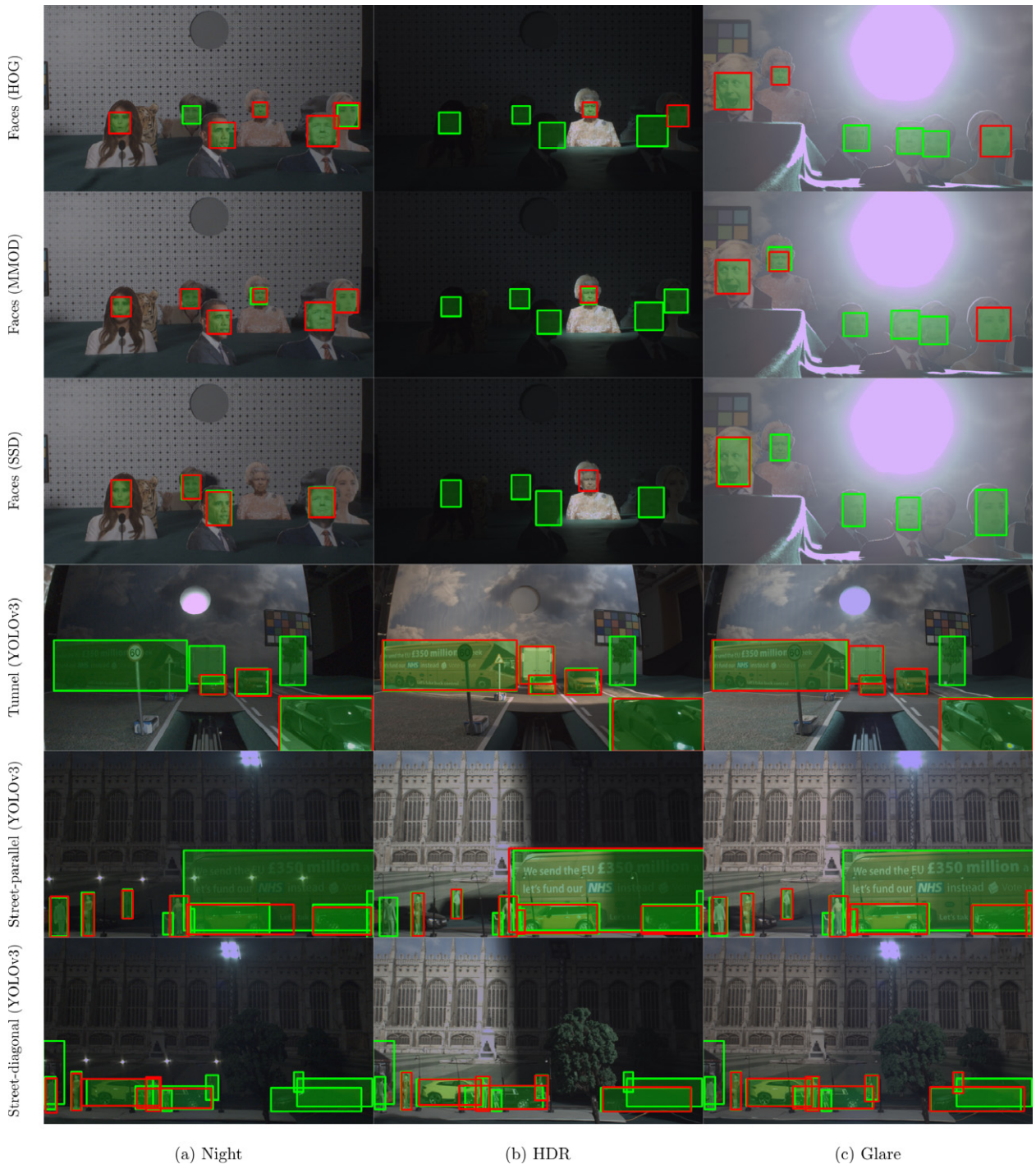


Figure 6. Bounding boxes for several face and object detection methods (rows) for three adversarial illumination conditions (columns), each *gamma* encoded (with $\gamma = 2.2$). Faces and objects detected by the various methods for the reference uniform illumination condition are shown as shaded green rectangles, while the results of the tested methods for other illumination conditions are shown as red rectangles.

Saturation of the pixels was unavoidable for the glare conditions, containing bright light sources. However, we selected exposures in which most objects were visible and not saturated.

The HDR image stacks were merged using HDRutils [31], which reduces estimation error in the presence of noise. We used 13 exposures, from 0.2 ms to 819.2 ms, separated by 1 stop. Demosaicing and white balance were performed after merging RGGB sub-pixels. The multi-exposure merging procedure resulted in HDR images with a minimal amount of noise and without any saturated regions. Such images can be used to simulate other cameras with different noise characteristics [29–31].

Table II. The performance of popular methods under different illumination conditions, averaged over 100 images in each sequence. All input images were encoded using a *gamma* encoding ($\gamma = 2.2$). The reference labels for each application were obtained from uniform conditions. (*) The lens and setup for images containing glare in the *Faces* sequence differed from other conditions. As a result, a different setup for the reference uniform condition was used.

Application	Method	Faces			Tunnel			Street-parallel			Street-diagonal		
		Night	HDR	Glare*	Night	HDR	Glare	Night	HDR	Glare	Night	HDR	Glare
Face detection (mIoU) ↑	HoG	0.75	0.15	0.66	✗	✗	✗	✗	✗	✗	✗	✗	✗
	MMOD	0.93	0.16	0.44	✗	✗	✗	✗	✗	✗	✗	✗	✗
	SSD	0.56	0	0.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
Object detection (mIoU) ↑	YOLOv3	0.92	0.63	0.76	0.52	0.85	0.9	0.71	0.76	0.88	0.78	0.82	0.9
Optical flow (EE) ↓	PolyExp	0.97	5.91	3.1	2.52	2.78	2.35	4.11	3.56	1.34	2.07	2.47	1.84
	Coarse2Fine	0.13	0.64	0.35	1.03	6.21	6	0.57	0.18	0.09	0.35	0.28	0.17
	RAFT	54.71	40.49	1.82	22.6	0.32	0.27	0.68	0.15	0.1	0.48	0.2	0.15

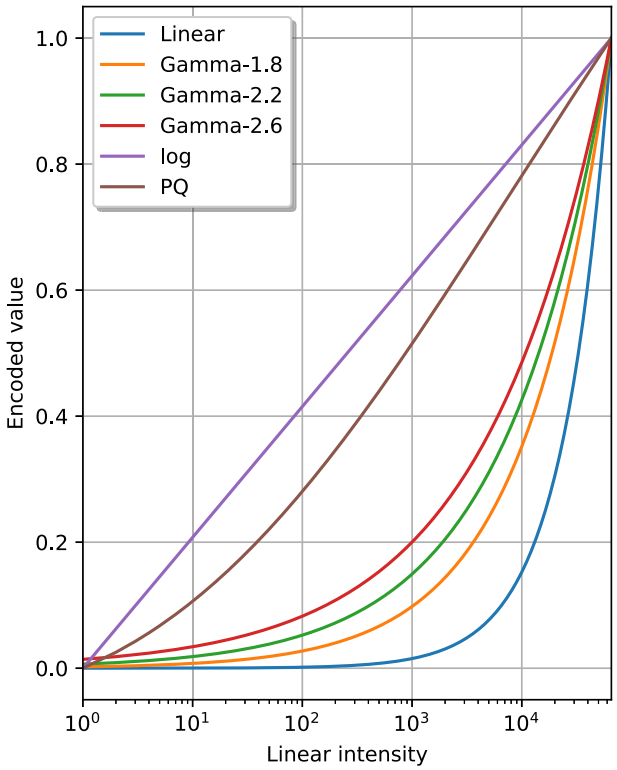


Figure 7. TFs used in our experiments to map linear color values to the encoded values, suitable for CV methods.

5.1 *Distribution of Pixel Values*

To visualize the differences between the illumination conditions, we plot histograms of pixel values in Figure 4 for each scene. It should be noted that the pixel values under uniform illumination are in the upper part of the dynamic range, making those images well-exposed. The HDR condition resulted in the widest histograms, often with two distinct modes. The histograms under the night illumination are shifted to the left, making images darker and more affected by noise. And finally, the histograms are shifted to the right in the glare condition due to the scattered light.

6. PERFORMANCE OF COMPUTER VISION METHODS

To demonstrate the usefulness of the new dataset, we perform two experiments: first, we test how much the performance of selected CV methods degrades due to challenging illumination and then test whether their performance can be improved when different TFs are used to encode frames. A suitable TF maps linear RGB pixel values to non-linear values, which can be represented at lower bit depth and tend to be more perceptually uniform. Note that we avoid labeling these methods *tone mapping* operators, as tone mapping is typically used to produce visually pleasing images for human consumption rather than machine vision. Our experiments form a preliminary study intended to confirm the selection of the adversarial illumination conditions and are not meant to be a comprehensive evaluation of all possible TFs. Such a larger experiment is planned for future work.

Traditional image signal processing pipelines are designed for best visual quality and are not optimized for CV algorithms [33, 34]. Many steps in the ISP pipeline may be redundant for some CV algorithms, and may even degrade their performance. Previous work [33] has shown that only two stages of the traditional ISP pipeline are critical in terms of machine vision, namely demosaicing and gamma encoding (or gamma compression). We follow this observation and simulate a simplified camera pipeline, shown in Figure 5. The first two steps of this pipeline, demosaicing and white balance, were explained in Section 5. The last two steps, encoding with a TF and quantization, were different for each experiment and tested method, and are explained in more detail in the following sections.

Our experiments were performed on publicly available implementations for three CV applications—optical flow, object detection, and face detection. We used the uniform condition encoded with the *gamma* TF (with an exponent of 1/2.2) to generate reference labels for evaluating methods. As described in Section 3.1.1, the uniform condition consists of well-exposed frames with low levels of noise. We created individual labels for each tested CV method and thus each

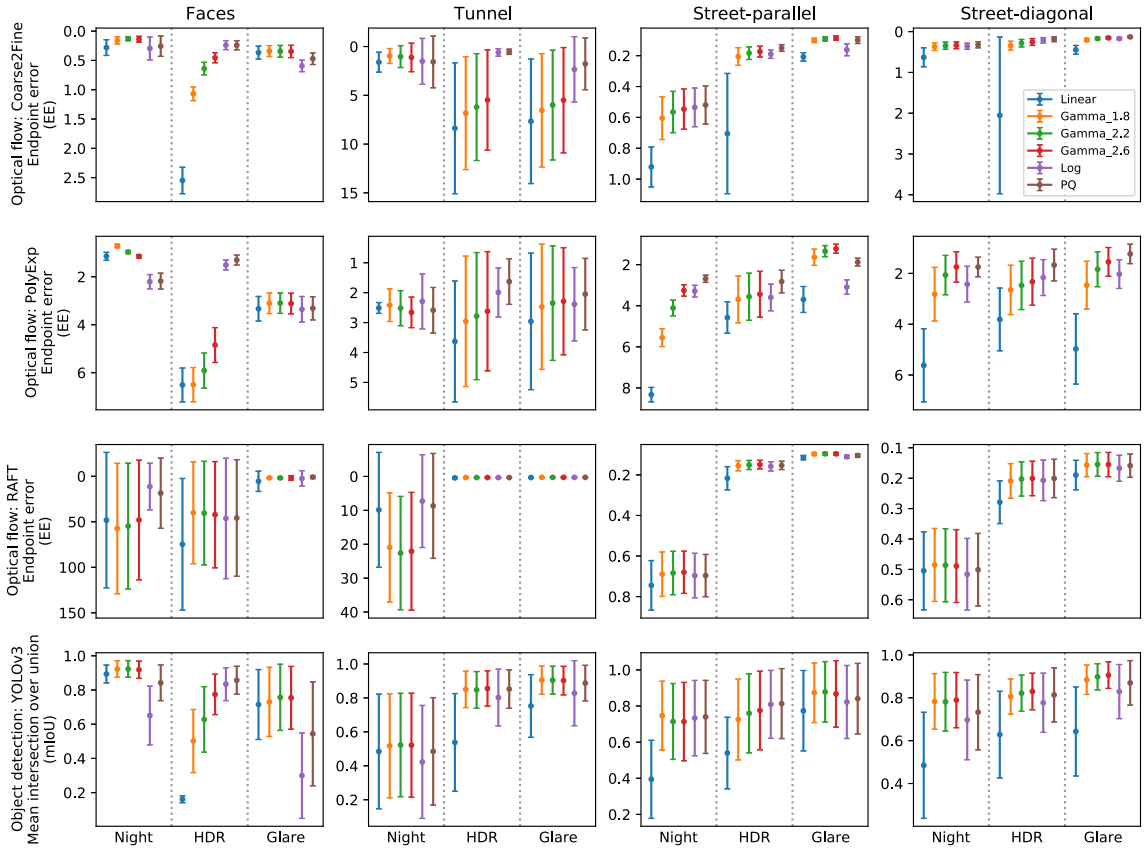


Figure 8. The performance of optical flow (top three rows) and object detection (bottom row) for all four scenes (columns) and three adversarial illumination conditions (three sections in each plot). Dots represent the mean, and the error bars represent the standard deviation for a sample of 100 frames. The y-axis is reversed for optical flow so that the points located higher on the plot correspond to the better performance.

face detection and optical flow method had its own set of reference labels. We inspected the reference labels and found that almost all objects were correctly detected in the uniform condition and therefore no manual labeling was necessary. It should be noted, however, that the reference condition should not be considered as the ground truth and is only meant to show the relative degradation in the performance due to the adversarial illumination.

We used dense optical implementations of a polynomial expansion algorithm [5], a coarse-to-fine algorithm [6], and a state-of-the-art deep neural network RAFT [35]. To compare predicted pixel-wise optical flow field $V_{\text{pred}}(p)$ to the reference $V_{\text{ref}}(p)$, we used the endpoint error (EE) metric

$$\text{EE} = \frac{1}{N}\sum_{p=1}^{N} |V_{\text{pred}}(p) - V_{\text{ref}}(p)|, \tag{1}$$

where N is the number of pixels.

For object detection we used a pretrained YOLOv3 [36] network. For face detection we used a pretrained single-shot detector (SSD) [37], face detector based on histogram of oriented descriptors (HOG) [38], and maximum margin object detection with CNN features (MMOD) [39]. Bounding boxes of both detection tasks were evaluated using the mean intersection over union (mIoU) metric

$$\text{mIoU} = \frac{area(R_{\text{pred}} \bigcap R_{\text{ref}})}{area(R_{\text{pred}} \bigcup R_{\text{ref}})}, \tag{2}$$

where $R_{\text{pred}} = \bigcup_{i=1}^{N} B_{\text{pred}}(i)$ is the union of all N bounding boxes predicted, and $R_{\text{ref}} = \bigcup_{j=1}^{M} B_{\text{ref}}(j)$ is the union of all M reference bounding boxes.

6.1 *Effect of Changing Illumination*

All images I_{in} were encoded with the *gamma* encoding, $I_{\text{out}} = I_{\text{in}}^{1/\gamma}$, where $\gamma = 2.2$. Such a gamma encoding is the most widely used TF in CV cameras, which produces images that can be directly viewed on standard monitors, as it approximates the TF used in the sRGB color space. Here, we test how the performance of the CV methods is degraded in adversarial illumination conditions.

The results of this experiment are listed in Table II. We observe a considerable difference in the results of the tested methods with changing illumination conditions. Figure 6 shows a qualitative comparison for the different conditions for the face and object detection methods.

The low scores of all face detection methods for the HDR illumination (see Table II) can be explained by the first three images of the second column of Fig. 6. The inefficient encoding of the input images results in 5 out of 6 faces

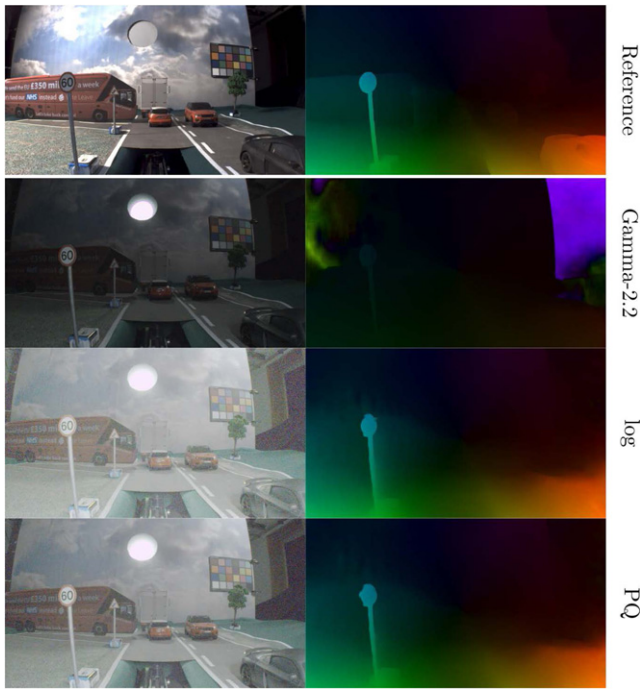


Figure 9. Failure of *gamma* encoding for night condition of the *Tunnel* scene when the optical flow is estimated by RAFT. The purple region in optical flow visualization for *gamma-2.2* indicates high velocity in the south direction, which is most likely caused by the dark region in that corner of the frame. The HDR TFs (*log* and *PQ*) can better handle this sequence. This is also the cause of the large error bars in some plots in Fig. 8.

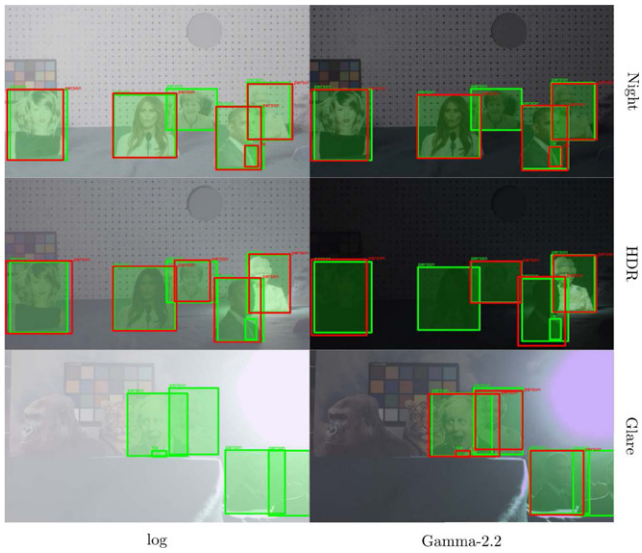


Figure 10. Object detection using YOLOv3 for images captured under different illumination conditions (rows) and encoded with *log* (left column) and *gamma* (right column) TFs. The *log* function produces brighter images by devoting more bits to dark regions as indicated by the left plot in Fig. 7. This leads to contrast reduction in bright regions posing problems to some object and face detectors for the Glare condition, but this is an effective strategy for other conditions.

being underexposed. In the glare condition (last column of Fig. 6), the faces present under the bright light source are not detected by any of the methods. The MMOD and HOG detectors are slightly more robust to glare and are able to detect faces further away from the source of glare. This explains their better performance for the same inputs.

Glare is not as much of a problem in the other scenes (*Tunnel* and *Street*) because the objects of interest are further away from the bright light source. In general, the methods produce lower mIoU scores for very dark objects in night illumination. This is likely a consequence of underexposed pixels rather than noise since underexposed objects pose a problem even in HDR conditions (for example, the bus and the truck in *Tunnel*, and the car and the person in the bottom-right shaded region in *Street-diagonal*).

6.2 *Comparison of TFs*

Next, we evaluate the performance of the CV methods when the input images are encoded using different TFs. These include a *linear* function (no TF, linear color values), *gamma* encoding (see Section 6.1) using three different exponents (1/1.8, 1/2.2, and 1/2.6), the *log* function, and the perceptual quantizer (*PQ*) TF, which is commonly used for encoding HDR content [40]. All the TFs are plotted using log-linear coordinates in Figure 7. The advantage of this visualization is that the slopes of the plotted functions correspond to the compression or expansion of contrast at a particular intensity value. The plots show that the TFs intended for HDR images (*log* and *PQ*) better preserve contrast across the entire range of values, while *gamma* and *linear* functions compress contrast at lower intensity values. The code for the transfer functions can be found at https://www.cl.cam.ac.uk/research/rainbow/projects/hdr4cv-dataset/.

Optical flow As shown in the top three rows of Figure 8, all tested optical flow methods show similar characteristics: they all benefit from HDR TFs (*log* and *PQ*) for most conditions. It should, however, be noted that there are also important differences between the methods. The Coarse2Fine method seems to be robust to noise, in the night condition but affected by the large dynamic range, in the HDR condition, especially for the *tunnel* scene (note the absolute values of the EE). PolyExp resulted in larger EE across all the conditions and is less robust to noise. Finally, RAFT resulted in huge EE for *Faces* and Tunnel. We investigated this issue and found that the cause was a dark region of the image, which captured a part of our laboratory outside the illuminated part of the scene. RAFT predicted very high velocity for this region, resulting in an EE an order of magnitude higher than for other methods (see Figure 9). For the two *Street* conditions, RAFT was robust in the HDR condition but was affected by noise in the night condition. We can conclude that the Coarse2Fine method is most robust to adversarial conditions and it should be combined with *PQ*.

Object detection The results for object detection (YOLOv3) shown in the bottom row of Fig. 8 are strongly

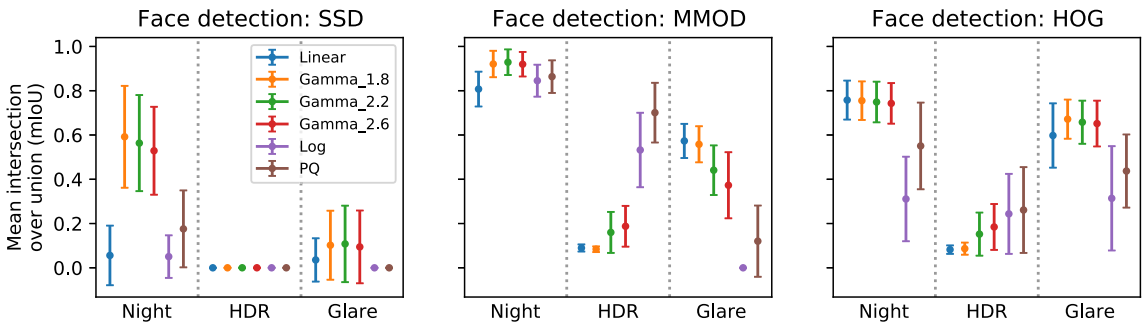


Figure 11. The performance of face detection methods for all transfer functions and all illumination conditions. Dots represent sample mIOU, and error bars represent the sample standard deviation over 100 frames. When some transfer functions are used, some methods are unable to detect even a single face. For example, notice the zero entries for the HDR and glare illuminations with the *log* function for SSD and MMOD.

affected by the adversarial illumination. The TFs intended for HDR data, *log*, and *PQ*, improve performance for HDR condition, but they also reduce performance for the two other conditions. This is because night and glare conditions contain objects of interest in the upper (brighter) part of the dynamic range. Since *gamma* encoding tends to enhance contrast in the upper part of the dynamic range, it achieves better performance for those two conditions but fails for HDR, where it is unable to reproduce contrast in the darker part of the frame, as shown in Figure 10.

Face detection The results plotted in Figure 11 show similar trends for all three face detection methods. As was the case with object detection, HDR TFs help in HDR conditions, but can also degrade performance for other conditions. For the same reason, *linear* color representation unexpectedly performs reasonably well for night and glare, though it still fails for the SSD method. It is also worth noting that the SSD face detector is more affected by the adversarial illumination than the two other methods. We can conclude, that face detection would benefit from an adaptive TF, which selectively reproduces contrast in the darker or brighter part of the dynamic range, depending on image content.

7. LIMITATIONS

Although our dataset was designed to be possibly realistic, it may not be suitable for applications that rely on accurate material reflectance properties, such as shape from shading or relighting. This is because many objects in our model scenes are cardboard cutouts and do not capture the richness of materials in the wild. The dataset could complement computer graphics datasets, which can potentially reproduce more accurate materials but may lack imperfections and artifacts caused by camera sensors (noise, quantization) and optics (glare).

8. CONCLUSIONS

We created a new dataset consisting of short video clips captured in adversarial illumination conditions. The main advantage of our dataset is that it includes uniformly illuminated scenes which result in images with minimal noise. With these, reference labels can be automatically generated for most CV methods. Such labels can be then used to measure the degradation in the performance of computer vision methods in adversarial conditions: night (high noise), HDR (high contrast), and glare (stray light in camera lens).

We used the dataset in our preliminary study to evaluate the robustness of popular methods for face detection, optical flow, and object detection under adversarial illumination conditions. We also studied whether the performance of these methods can be further improved under challenging lighting conditions by selecting an appropriate transfer function. The results suggest that as expected, the popular *gamma* encoding is unsuitable for HDR scenes. At the same time, the transfer functions intended for HDR scenes, such as *log* or *PQ*, reduce contrast in bright regions leading to rather poor performance in well-exposed images. We plan to expand this study to consider more advanced, adaptive transfer functions, which could further improve the robustness of CV methods under adversarial illumination is in preparation.

ACKNOWLEDGMENT

We would like to thank Maryam Azimi for the valuable comments.

REFERENCES

[1] L. Murmann, M. Gharbi, M. Aittala, and F. Durand, "A dataset of multi-illumination images in the wild," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2019), pp. 4080–4089.

[2] L. Yihang, B. Yan, L. Jun, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2019), pp. 3230–3238.

[3] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," IEEE Trans. Multimedia **20**, 2593–2604 (2018).

[4] A. Tsirikoglou, G. Eilertsen, and J. Unger, "A survey of image synthesis methods for visual machine learning," Computer Graphics Forum **39**, 426–451 (2020).

[5] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, edited by Josef Bigun and Tomas Gustavsson (Springer, Berlin Heidelberg, 2003), pp. 363–370.

[6] C. Liu, "*Beyond Pixels: Exploring New Representations and Applications for Motion Analysis.*" Ph.D. thesis (Massachusetts Institute of Technology, 2009).

[7] S. M. Safdarnejad, Y. Atoum, and X. Liu, "Temporally robust global motion compensation by keypoint-based congealing," in *Computer Vision—ECCV 2016*, edited by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Springer International Publishing, Cham, 2016), pp. 101–119.

[8] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *2003 Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2003), Vol. 1, pp. I–I.

[9] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," *German Conf. on Pattern Recognition* (Springer, Berlin Heidelberg, 2014), pp. 31–42.

[10] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," Int. J. Comput. Vis. **92**, 1–31 (2011).

[11] L. Tsung-Yi, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European Conf. on Computer Vision* (Springer, Berlin, Heidelberg, 2014), pp. 740–755.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." Technical Report 07-49, University of Massachusetts, Amherst (2007).

[13] L. Ziwei, L. Ping, W. Xiaogang, and X. Tang, "Deep learning face attributes in the wild," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2015), pp. 3730–3738.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 4401–4410.

[15] S. Martin, K. Yuen, and Mohan M Trivedi, "Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge," *2016 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, Piscataway, NJ, 2016), pp. 1010–1014.

[16] A. Mohan, R. Bailey, J. Waite, J. Tumblin, C. Grimm, and B. Bodenheimer, "Tabletop computed lighting for practical digital photography," IEEE Trans. Visualization and Computer Graphics **13**, 652–662 (2007).

[17] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2018), pp. 1692–1700.

[18] Y. Atoum, "*Detecting Objects under Challenging Illumination Conditions.*" Ph.D. thesis (Michigan State University, 2018).

[19] H. Alismail, B. Browning, and S. Lucey, "Robust tracking in low light and sudden illumination changes," *2016 Fourth Int'l. Conf. on 3D Vision (3DV)* (IEEE, Piscataway, NJ, 2016), pp. 389–398.

[20] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi, "Deep image-based relighting from optimal sparse samples," ACM Trans. Graphics **37**, 1–13 (2018).

[21] M. E. Helou, R. Zhou, S. Süsstrunk, R. Timofte, M. Afifi, M. S. Brown, K. Xu, H. Cai, Y. Liu, L.-W. Wang, Z.-S. Liu, C.-T. Li, S. D. Das, N. A. Shah, A. Jassal, T. Zhao, S. Zhao, S. Nathan, M. P. Beham, R. Suganya, Q. Wang, Z. Hu, X. Huang, Y. Li, M. Suin, K. Purohit, A. N. Rajagopalan, D. Puthussery, P. S. Hrishikesh, M. Kuriakose, C. V. Jiji, Yu Zhu, L. Dong, Z. Jiang, C. Li, C. Leng, and J. Cheng, "AIM 2020: scene relighting and illumination estimation challenge," in *Computer Vision – ECCV 2020 Workshops*, edited by A. Bartoli and A. Fusiello (Springer International Publishing, Cham, 2020), pp. 499–518.

[22] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. Debevec, and R. Ramamoorthi, "Single image portrait relighting," ACM Trans. Graph. **38**, 79, 1–12 (2019).

[23] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 4340–4349.

[24] Y. Cabon, N. Murray, and M. Humenberger, Virtual KITTI 2. Preprint arXiv:2001.10773, (2020).

[25] Z. Sicheng, Y. Xiangyu, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. on Neural Networks and Learning Systems* (IEEE, Piscataway, NJ, 2020), pp. 1–21.

[26] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2017), pp. 2962–2971.

[27] M. Callegari, Q light controller+. https://www.qlcplus.org, (2020).

[28] B. Muthukadan, Selenium with python. https://selenium-python.readthedocs.io (2020).

[29] A. Abdelhamed, M. Brubaker, and M. Brown, "Noise flow: noise modeling with conditional normalizing flows," *2019 IEEE/CVF Int'l. Conf. on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2019), pp. 3165–3173.

[30] C. Aguerrebere, J. Delon, Y. Gousseau, and P. Musé, "Study of the digital camera acquisition process and statistical modeling of the sensor raw data." Technical Report, (2013).

[31] P. Hanji, F. Zhong, and R. K. Mantiuk, "Noise-aware merging of high dynamic range image stacks without camera calibration," *Advances in Image Manipulation (ECCV workshop)* (Springer, Berlin Heidelberg, 2020), pp. 376–391.

[32] D. Menon, S. Andriani, and G. Calvagno, "Demosaicing with directional filtering and a posteriori decision," IEEE Trans. Image Process. **16**, 132–141 (2006).

[33] M. Buckler, S. Jayasuriya, and A. Sampson, "Reconfiguring the imaging pipeline for computer vision," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2017), pp. 975–984.

[34] R. M. H. Nguyen and M. S. Brown, "Why you should forget luminance conversion and do something better," *CVPR* (IEEE, Piscataway, NJ, 2017), pp. 5920–5928.

[35] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," *European Conf. Computer Vision* (Springer, Berlin Heidelberg, 2020), pp. 402–419.

[36] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement." Preprint arXiv:1804.02767 (2018).

[37] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *European Conf. on Computer Vision* (Springer, Berlin, Heidelberg, 2016), pp. 21–37.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE, Piscataway, NJ, 2005), Vol. 1, pp. 886–893.

[39] D. E. King, "Max-margin object detection." Preprint arXiv:1502.00046, (2015).

[40] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual signal coding for more efficient usage of bit codes," SMPTE Motion Imaging J. **122**, 52–59 (2013).