IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 7, 2020, accepted July 3, 2020, date of publication July 20, 2020, date of current version August 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010340

Backward Compatible Object Detection Using HDR Image Content

RATNAJIT MUKHERJEE[1,2], **MIGUEL MELO**[1], **VÍTOR FILIPE**[1,3], **ALAN CHALMERS**[4], **AND MAXIMINO BESSA**[1,4]

[1]INESC TEC, 4200-465 Porto, Portugal
[2]Advanced Research Laboratory, NavInfo Europe B.V., 5657DB Eindhoven, The Netherlands
[3]Universidade de Trás-os-Montes e Alto Douro, 5001-801 Vila Real, Portugal
[4]WMG, University of Warwick, Coventry CV4 7AL, U.K.

Corresponding author: Maximino Bessa (maxbessa@utad.pt)

This work was supported by the U.S. Office of Naval Research, through the mechanism for Naval International Cooperative Opportunities (NICOP), under the Grant N629091612243.

ABSTRACT Convolution Neural Network (CNN)-based object detection models have achieved unprecedented accuracy in challenging detection tasks. However, existing detection models (detection heads) trained on 8-bits/pixel/channel low dynamic range (LDR) images are unable to detect relevant objects under lighting conditions where a portion of the image is either under-exposed or over-exposed. Although this issue can be addressed by introducing High Dynamic Range (HDR) content and training existing detection heads on HDR content, there are several major challenges, such as the lack of real-life annotated HDR dataset(s) and extensive computational resources required for training and the hyper-parameter search. In this paper, we introduce an alternative backwards-compatible methodology to detect objects in challenging lighting conditions using existing CNN-based detection heads. This approach facilitates the use of HDR imaging without the immediate need for creating annotated HDR datasets and the associated expensive retraining procedure. The proposed approach uses HDR imaging to capture relevant details in high contrast scenarios. Subsequently, the scene dynamic range and wider colour gamut are compressed using HDR to LDR mapping techniques such that the salient highlight, shadow, and chroma details are preserved. The mapped LDR image can then be used by existing pre-trained models to extract relevant features required to detect objects in both the under-exposed and over-exposed regions of a scene. In addition, we also conduct an evaluation to study the feasibility of using existing HDR to LDR mapping techniques with existing detection heads trained on standard detection datasets such as PASCAL VOC and MSCOCO. Results show that the images obtained from the mapping techniques are suitable for object detection, and some of them can significantly outperform traditional LDR images.

INDEX TERMS High dynamic range (HDR), low dynamic range (LDR), object detection, faster RCNN, SSD, R-FCN.

I. INTRODUCTION

The introduction of Convolution Neural Networks (CNN) has brought about a complete paradigm shift in object recognition and detection, which has been a major challenge in computer vision [1]. State-of-the-art CNN based object detectors [2]–[5] have been able to achieve unprecedented accuracy in generic object detection tasks, for example, $\geq$ 80% accuracy on the Pascal Visual Object Challenge (PASCAL VOC) and $\geq$ 50% in the challenging Microsoft Common Object in Context (MS COCO) detection track. However, in spite of this unprecedented accuracy, a common issue with most existing detection models is their inability to accurately detect salient objects in challenging and/or extreme lighting conditions. This can be attributed to the fact that existing detection models are typically trained on generic object detection datasets comprising mostly of well-exposed and/or moderately lit 8/bits/pixel/channel Low Dynamic Range (LDR) images. In these images, a large portion of the scene information is captured as midtones with little or no information in the under-exposed or over-exposed regions. Thus current detection models trained on existing datasets are unable to extract salient features from scenes with challenging lighting conditions that are typically encountered in real-world scenarios.

One of the solutions to address this problem is to introduce High Dynamic Range (HDR) and Wide colour Gamut (WCG) imaging which can capture the entire range of lighting

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

conditions and colour gamut seen by the human eye [6]. Subsequently, the captured HDR and WCG content can be used to create annotated datasets and train existing detection heads. Unlike the traditional LDR image trained models, HDR trained models would thus be able to take advantage of the extended dynamic range and colour gamut to extract salient information from both the under-exposed and over-exposed regions of a scene in addition to the mid-tones. However, there are a large number of challenges which need to be overcome before retraining of detectors on HDR datasets is possible. Firstly, the amount of unique and true HDR content (available for annotation) is remarkably low. Secondly, there are no existing annotation tools which support native HDR image content. Thirdly, due to the intrinsic differences in source content, retraining and hyper-parameter search for state-of-the-art detectors is computationally time and energy consuming. Finally, due to the lack of diverse true HDR content, retrained detectors would be unable to generalise well to *out-of-distribution* real-world data.

Given these challenges, in this paper, we propose a robust methodology for object detection in extreme and/or challenging lighting conditions while ensuring backward compatibility with most current detection models trained on existing LDR datasets. The proposed methodology explores several techniques to transform native HDR content to LDR content using generic transfer functions (explained later in section II-A) such that the resultant 8-bits/pixel/channel LDR images are able to reproduce salient scene information (both luminance and chroma) by essentially compressing and faithfully reproducing the tones of native HDR content. Additionally, we also conduct a comprehensive evaluation, with an out of distribution dataset (OOD), to study the effects of seven different HDR to LDR mapping techniques, compare them with traditional LDR approaches on three CNN based detection models, and measure their detection accuracy.

II. RELATED WORK

As the contributions of this paper encompass prior research from two significantly different research areas, a detailed description of all relevant previous work from both research areas is out of scope and hence in this section we only provide a brief overview of the prior research directly used in this work.

A. HDR TO LDR MAPPING

The vast majority of the imaging devices are only able to capture, process, and display LDR content with ITU-R BT.709 colour gamut [7], i.e., they are unable to represent the full colour gamut and scene dynamic range as seen by the human eye. However, true HDR content is able to capture, store and process more than 16-stops of scene dynamic range with either ITU-R BT.709 [7] or ITU-R BT.2020 [8] to fully encompass the dynamic range that the human eye can see with minimal eye adaptation [6]. Also, unlike WCG content, typically captured using the BT 2020 colour space [8], true HDR content (16-bit floating point format) cannot be processed and displayed using hardware-based LDR processing and display devices. To represent HDR content on LDR devices, the dynamic range and colour gamut information (not necessarily WCG) of native HDR content needs to be compressed using transfer functions. These are designed to compress maximal luminance information especially from the over- and under-exposed regions of a scene while preserving the chroma information. Such transfer function based algorithms are typically known as tone-mapping operators (TMOs), image-appearance/colour appearance based operators, or exposure fusion operators. Although, a few Field Programmable Gate Array (FPGA)-based real-time HDR to LDR processing devices exist [9], they typically use scaled-down versions of perceptual and global TMOs which by their intrinsic simplicity are unable to faithfully represent the complexity of a scene's luminance and chroma information. Given these challenges, the following transfer functions have been considered as a part of this work.[1] Figure 1 provides a visual representation of each HDR to LDR transfer function and the justification for the selection of these transfer functions is given later in section III-D.

1) PHOTOGRAPHIC TONE MAPPING OPERATOR ((ReinhardTMO)

Introduced in 2002 by Reinhard *et al.* [10], this HDR to LDR mapping technique leverages the time tested photographic tone-compression technique (Zonal System) first proposed by Ansel Adams [11]. The compression process first maps the image by scaling the original scene luminance using the log-average luminance as the approximation to the key of the scene. Subsequently, it focuses on photographic ''dodging and burning'' principles which change the exposures of different parts of the image such that the darker and brighter regions of a scene can be faithfully reproduced.

2) DISPLAY ADAPTIVE TONE MAPPING OPERATOR (MantiukTMO)

This TMO, proposed in 2008 by Mantiuk *et al.* [12], allows the configuring of different parameters to take into consideration the display features and the environment where the content is being viewed in order to optimise its output. Besides the possibility of fine-tuning each parameter individually, the TMO includes a preset of different profiles to target different viewing conditions (display size, ambient luminance, etc.). The tone-mapping process starts by processing the image, based on a display model and viewing conditions. Subsequently, the image is processed by taking into account human visual models having as basis the tone-reproduction for realistic computer-generated images [13] and the adaptation for realistic image display [14].

[1] The transfer functions chosen for this work, typically map 16-bits/pixel/channel floating point HDR content captured using the BT709 colour space to 8-bits/pixel/channel sRGB colour space.

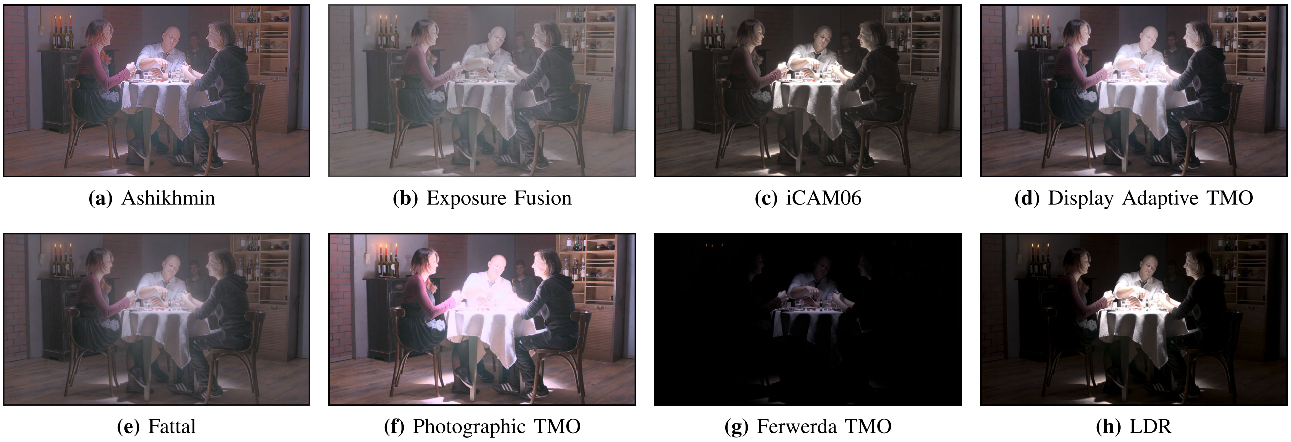


FIGURE 1. Visual representation of seven HDR to LDR mapping techniques along with an LDR representation.

3) ADAPTIVE LOGARITHMIC MAPPING (FattalTMO)

Based on the assumption that the human eye is more sensitive to local intensity ratio changes rather than absolute luminances, Fattal *et al.* [15] proposed a tone-compression technique based on the gradient domain compression. This TMO identifies large gradients at different scales and attenuates their magnitudes while keeping their direction unaltered. All the computations are performed in the logarithmic scale of luminances, where higher significant gradients are penalised more to compress drastic luminance changes while preserving detail. The attenuation of the gradients is achieved by applying an appropriate spatially variant attenuation mapping to the magnitudes of the derivatives at each pixel. The progressiveness of the method is performed by constructing a Gaussian pyramid to achieve a multi-resolution edge detection scheme. To handle the extensive system of linear equations, the authors propose using the Full Multigrid Algorithm [16], with Gauss-Seidel smoothing iterations.

4) TONE-MAPPING ALGORITHM FOR HIGH CONTRAST IMAGES (AshikhminTMO)

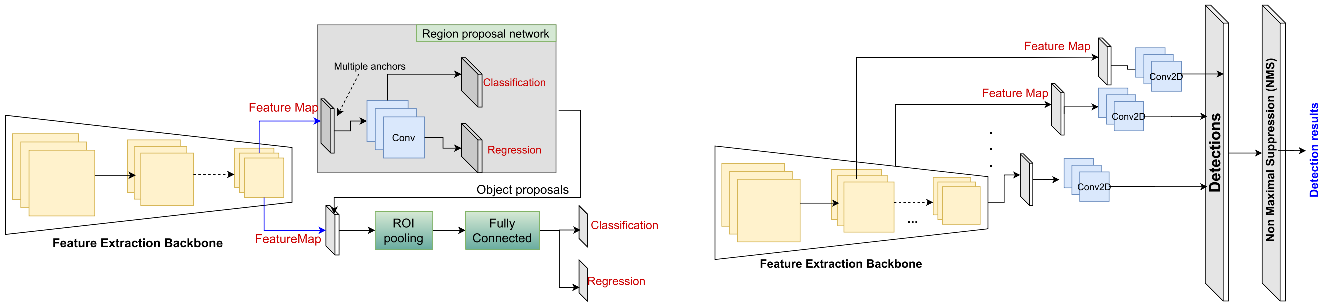
This approach takes advantage of knowledge of the human visual system (HVS) and consists of three main steps: local luminance adaptation, compression of image values to display values, and a final pass for detail re-introduction that could have been lost in the previous step. For estimating the local adaptation level, the TMO balances the local contrast signal whilst maintaining information about image details based on the average luminance over a pixel neighbourhood. The compression of image values to display values is achieved by employing *threshold vs intensity* functions having as reference the world luminance. As this step works at a contrast level, some detail can be lost. Thus, the third step is performed to recover detail by multiplying the obtained image by a detailed image that is given by the ratio of pixel luminance to the corresponding local world adaptation level.

5) VISUAL ADAPTATION MODEL (Ferwerda TMO)

Introduced by Ferwerda *et al.* [17], this HDR to LDR mapping technique was developed for targeting realistic image synthesis based on the physical features of the HVS (i.e. scotopic, mesopic and photopic vision) and was based on previous work regarding both brightness-based operators [18] and contrast based operators [19]. The first step of the TMO consists of calculating the luminance adaptation of the input image. Then, the scotopic and photopic vision scaling factors are calculated by employing *threshold vs intensity* functions. For the mesopic values (between scotopic and photopic conditions), a step is taken to combine the scotopic and photopic luminance values according to a constant k that works as a scale factor that ranges from 0 to 1 according to the scotopic adaptation level. The final step of the TMO is to normalise the output values taking into account the maximum luminance of the display device.

6) EXPOSURE FUSION

Unlike the previously mentioned HDR to LDR mapping techniques, this method [20] combines a bracketed exposure sequence which would generate an HDR image, into an LDR image without actually generating the HDR image. This simplifies the processing pipeline when the goal is to directly obtain the compressed LDR image as it saves all the processing needed for the HDR generation. To achieve the dynamic range compression, the method selects the richest content of each exposure based on a scalar-valued weight map. To determine such regions, the contrast, saturation and well-exposedness of the image are considered. The contrast is evaluated by using Laplacian filters to detect edges and texture; the saturation is evaluated by computing the standard deviation of the RGB channels; and the well-exposedness considers the raw intensities of each image channel that are close to 0.5 after applying a Gauss curve. Each parameter generates a weight value that is used to fuse the image sequence. To avoid the image having artefacts introduced



(a) Schematic diagram of Region Proposal-based (two-stage) object detection algorithms

(b) Schematic diagram of unified (one-stage) based object detection algorithms

FIGURE 2. Schematic diagram of the two main approaches to object detection. Figure 2a demonstrates the region proposal based approach while Figure 2b demonstrates the unified framework approach.

by the fact that the different regions have different absolute intensities, a multi-resolution blending based on Laplacian decomposition is applied.

7) IMAGE APPEARANCE MODEL OF HDR RENDERING (iCAM06)

Kuang *et al.* [21] proposed a new image appearance model, designated iCAM06, designed specifically for HDR image rendering. Based on the iCAM framework [22], the new model incorporates the spatial processing models in the HVS for contrast enhancement and photo-receptor light adaptation functions that enhance local details in highlights and shadows. The original HDR image is first converted to the CIE-XYZ [23] colour space and subsequently decomposed into a base and a detail layer wherein the base layer is obtained using an edge-preserving bilateral filter [24] and the detail layer is obtained by subtracting the base layer from the original image. The base layer first undergoes chromatic adaptation which is achieved by converting the CIE-XYZ image to a spectrally sharpened RGB image using the MCAT02 transformation matrix [22]. The incomplete adaptation factor is computed as a function of adaptation luminance and the surround factor. Subsequently, the spectrally sharpened RGB image is converted from the CAT02 space to the Hunt–Pointer–Estevez fundamentals which is where the resultant RGB signal undergoes a non-linear tone compression using a non-linear response function for both rods and cones derived from the Hunt Model [25]. The tone-compressed RGB image is then converted to the perceptually uniform colour opponent space IPT [26], which is desired because image attribute adjustments do not affect other attributes. To preserve the naturalness of the rendered tone-compressed image, the detail layer is enhanced to predict the Stevens effect and the P&T channels of the base layer are enhanced to predict the Hunt effect [25]. Finally, the enhanced base and detail layers are combined to produce an enhanced perceptually uniform output image. This is displayed on the target device by converting the IPT image to an RGB signal followed by an inverse chromatic adaptation.

B. CNN BASED OBJECT DETECTION

State-of-the-art CNN based object detection algorithms may be classified into two categories i.e. a) *two-stage* or regional-proposal based algorithms and b) *single stage* or unified framework algorithms, as illustrated in Figure 2.

1) TWO-STAGE OBJECT DETECTION

Also known as *Region proposal* based detectors, these methods first generate category agnostic region proposals from the entire image followed by feature extraction from the proposed regions. Subsequently, the proposed regions are indexed and fed to two network heads composed of fully connected layers. The first (classification) head consists of a softmax layer which determines the object category *'C'* (labels) along with confidence scores *'S'* $\in [0, 1]$. The second (regression) head is a bounding box regressor which determines the spatial location *'BBox'* of the object of interest. Finally, the indexes are matched and the corresponding labels, confidence scores, and bounding boxes are passed to a threshold algorithm known as *non-maximal suppression* (NMS). This discards the model predictions ('C', 'S' and 'BBox') with lower confidence scores than the chosen value, typically $\geq>$ 0.5. Noteworthy examples of *regional proposal* algorithms include the R-CNN family of object detectors [2], [27]–[29] and R-FCN [30].

2) ONE-STAGE OBJECT DETECTION

Also known as *unified framework*, this approach consists of a single *end-to-end* feed forward network performing classification and regression in a monolithic setting that does not require regional proposal or post classification. This design encapsulates all computations in a single feed-forward network which can be trained and optimised end-to-end, thereby reducing inference time significantly, making them ideal for real-time detection purposes. Important examples of this approach include, but not limited to, You Only Look Once (YOLO) [31]–[33], Single Shot Multibox Detector (SSD) [3], and the recently proposed paired keypoint based detection algorithm, such as Cornernet [4] and more recently, Objects as Points [5].

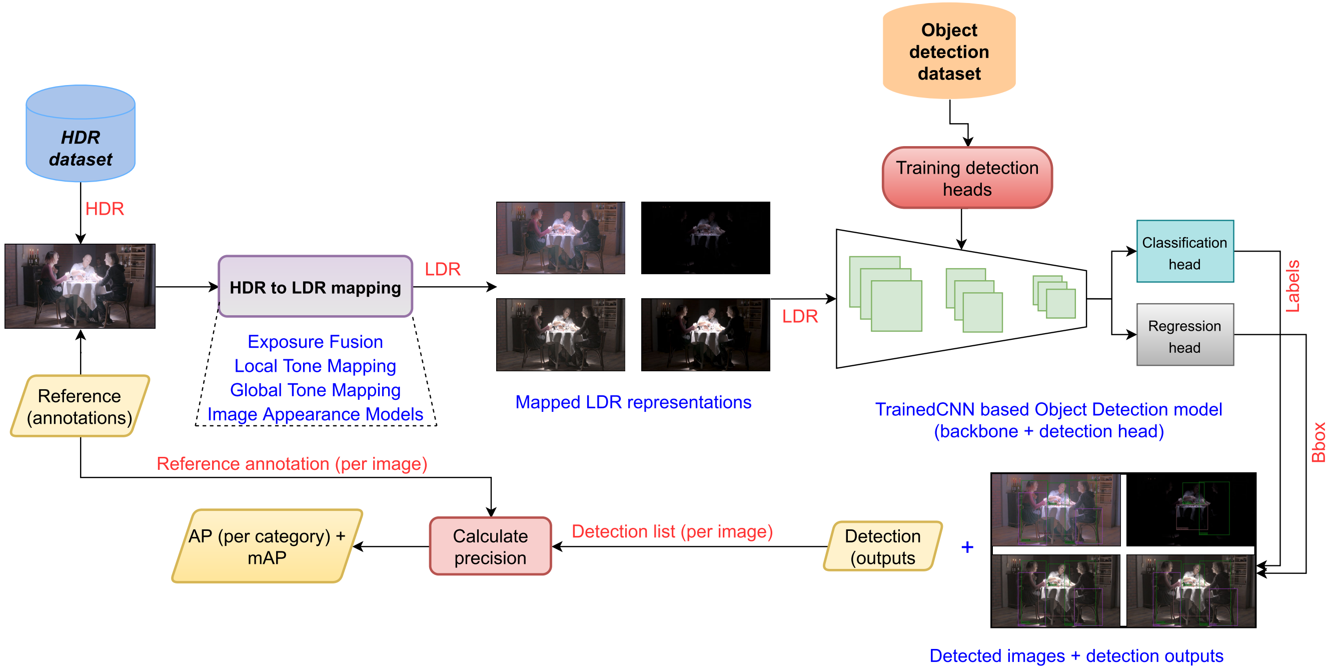


FIGURE 3. Schematic diagram of the proposed methodology for backward compatible HDR object detection.HDR/WCG 16-bits/pixel/channel images are fed to a selected HDR to LDR mapping technique. The resultant 8-bits/pixel/channel LDR images (with sRGB colour gamut) are then passed to the pre-trained detection model for inference. Final predictions for each image are matched with ground-truth annotations to calculate the AP per category (see section III for details).

Excellent detailed surveys on the evolution of CNN based object detection algorithms are given in [34]–[36]. Figures 2a and 2b provide a simplified schematic representation of the two different categories of CNN based object detection algorithms.

In section III-B, we discuss three of the above mentioned detectors which were chosen for the current work and provide the necessary justification for the choice.

III. METHODOLOGY

In this section, we describe in detail, the proposed methodology to use existing CNN based detection heads for object detection in HDR image/video content. This is achieved without the expensive creation of large annotated HDR datasets and computationally extensive training and hyper-parameter search for the optimal performance of existing detectors on the same HDR datasets. For the sake of brevity, we organised this section into the following subsections, each describing stages of the evaluation process. An overall schematic diagram of the evaluation process is given in Figure 3.

A. CHOICE OF OBJECT DETECTION HEADS

In section II-B we outlined the two key techniques used in object detection, namely *region-proposal* based CNNs and *unified framework* based CNNs. From the *region-proposal* family of object detectors, we chose Faster R-CNN [2] and R-FCN [30]. This is because these two detection heads are some of the first and seminal works on end-to-end trainable detection heads from the R-CNN family and do not require multi-stage progressive training such as R-CNN [27] and Fast R-CNN [28]. Furthermore, prior research [29], [37], [38] has demonstrated that *region-proposal* based detection heads are typically more accurate than *unified framework* based detection heads.

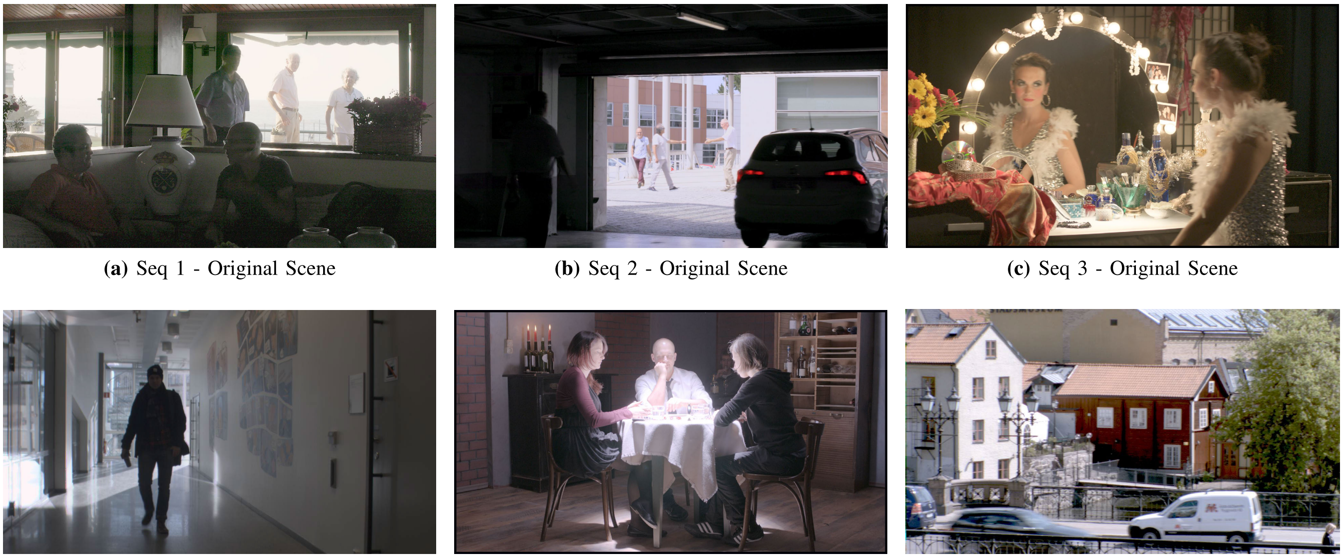
Amongst the *unified-framework* detection heads, two of the most widely used are YOLO [31] and SSD [3]. However, YOLO [31] is not particularly suitable for objects located in close spatial proximity and objects with a strong variance in size and aspect ratios as even the recent versions [32], [33] of the detection head are limited to three scales. On the other hand, SSD was specifically designed to predict object locations across various scales (≈ 6-7) and aspect ratios by appending additional convolution layers which gradually decrease in size and computing a fixed number of predictions in diverse aspect ratios resulting in prediction robustness across various scales and object sizes. Thus, our final choice was SSD.

To ensure the least amount of bias and variation, all three detection heads consisted of the same feature extraction backbone i.e. VGG 16 [39].

B. TRAINING OF OBJECT DETECTION HEADS

Firstly, to ensure the least amount of variation and jitter in training detection heads, the VGG 16 backbone (for all detection heads) was initialised with ImageNet [40] trained weights. Furthermore, convolution layers in the detection heads were initialised randomly using a fixed random seed value.

Amongst the most widely used detection LDR datasets, we chose the Pascal VOC dataset [41] which contains a total



(**a**) Seq 1 - Original Scene (**b**) Seq 2 - Original Scene (**c**) Seq 3 - Original Scene

(**d**) Seq 4 - Original Scene (**e**) Seq 5 - Original Scene (**f**) Seq 6 - Original Scene

FIGURE 4. Sample images from the OOD dataset. Images tone-mapped for representation. Average dynamic range is ≈ 18 stops.

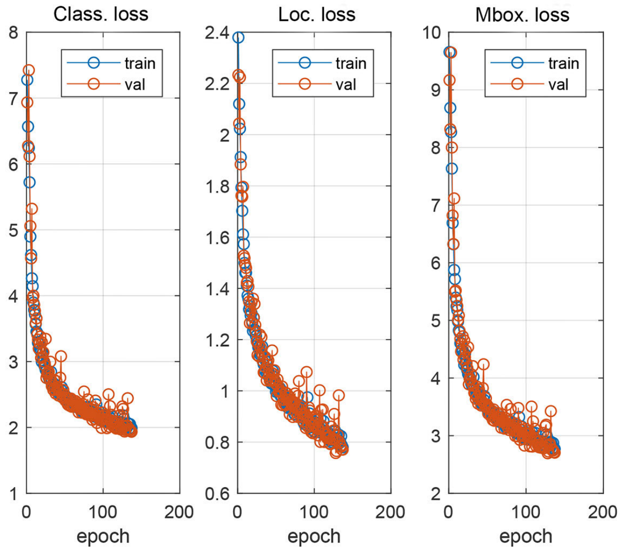


FIGURE 5. An example SSD training routine over 120 epochs showing train and validation loss for classification, localization (single box) and multi box.

of 20 object categories divided into two datasets VOC '07 and '12, respectively, with a combined total of 21,493 images containing 52,090 annotations. We combined training and validation sets of both VOC 2007 and 2012 for training and use the VOC 2007 test set for validation purposes (500 images at a time). The final test was conducted over the entire test set. To ensure the least amount of variation, we computed training loss and testing accuracy for a total 120 epochs, with a batch size of 32 on an Nvidia GTX 1080ti with 11 GB of virtual memory. The inference speed was also computed on the same GPU. The goal of the training procedure was to replicate the results provided in the original literature.

An example training routine of SSD over 120 epochs is given in Figure 5.

C. HDR DATASET CREATION

Unlike LDR, where a large number of computer vision datasets in various domains are available [41]–[44], the number of native HDR image/video datasets is extremely limited. A comprehensive search results in ≈ 3000-4000 native HDR images and ≈ 40-50 video sequences (with a duration of ≈ 10-15 seconds each). Out of the available datasets, only a fraction can be used for practical object detection purposes (containing more than one object).

Due to the limited availability of native HDR image/video material and expensive manual annotation procedure, we first shortlisted a total of ≈ 3500 images from multiple image datasets and 8 HDR video sequences. Subsequently, by means of exhaustive manual check, the shortlisted 3500 images were further pruned to a carefully curated set of 1289 images *(hereon referred to as the OOD dataset)* based on the following selection criterion: the dataset should contain images captured by a wide variety of medium to high dynamic range imaging devices such as Digital SLRs [45], Spheron VR [46], [47] and ARRI Alexa [48]) with an overall range of ≈ 17-21 stops.

The OOD dataset was subsequently downsampled to a resolution of 1920×1080 pixels and manually annotated by 4 annotators with an age range of $\approx 25 - 30$ years. The resultant dataset of 1289 images contains a total of 8638 annotations spanning six different object categories. The category-wise annotation histogram is given in Figure 6 and sample images from the OOD dataset are shown in Figure 4.

D. CHOICE OF HDR TO LDR MAPPING TECHNIQUES

To date, a large number of HDR to LDR mapping techniques have been proposed. An excellent review of these mapping techniques is available in [6]. However, as mentioned earlier in section II-A, the HDR mapping techniques are

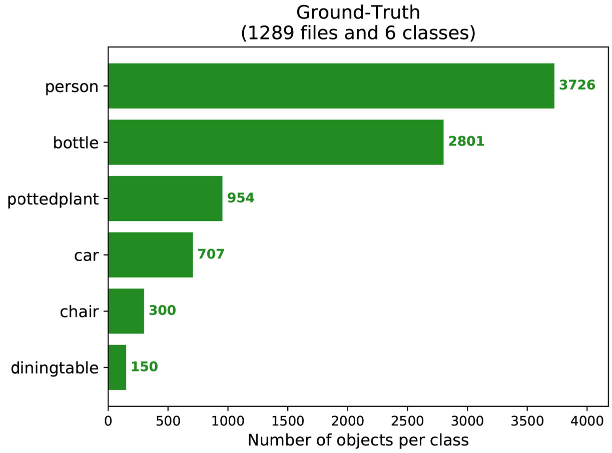


FIGURE 6. Salient features of the OOD dataset.

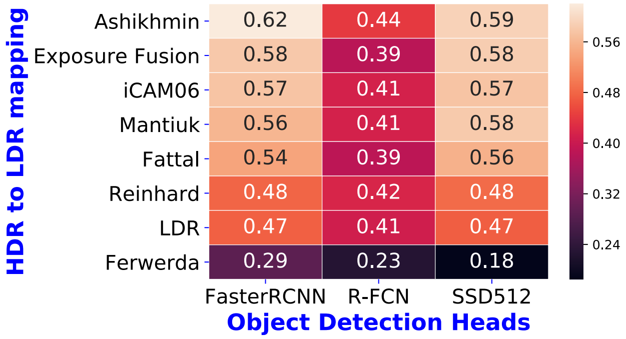


FIGURE 7. Heatmap of mean average precision - across all detection models, HDR mapping techniques and traditional LDR).

typically classified into a few categories, namely *global operators* [10], [17], local operators [15], [49], exposure fusion/blending [20], scene reproduction operators [12], and Image/colour Appearance models [21].

The primary selection criteria of HDR mapping techniques used in this work is that each of these mapping techniques are not only widely used, but also representative of a particular type of mapping technique. Furthermore, the selection was also inspired by prior research in this area [50]–[53] where the selected mapping techniques have outperformed other mapping techniques belonging to the same category.

In addition, to ensure the least amount of variation between the outputs of each mapping techniques, the usage parameters of each were set to the default values and fine-tuned in accordance with the prior research. The fine-tuning was carried out and validated by four experts (who each had at least 5 years' experience in HDR) to ensure which were the best settings across all TMOs. Finally, the pure LDR image (exposure) used in the selection, represents the middle (0^{th}) exposure of an HDR image based on the overall luminance of the scene.

E. EVALUATION PROCEDURE

As shown in Figure 3, the evaluation procedure consists of multiple steps which were conducted in parallel.

First, we trained the three detection heads using the PASCAL VOC dataset following the procedure mentioned in section III-B, resulting in three trained detection models. In parallel, the OOD dataset with 1289 images was created as mentioned in section III-C. Finally, we integrated the evaluation pipeline. During the evaluation process, HDR images are fed sequentially to each one of the seven HDR to LDR mapping techniques. The resultant mapped LDRs are then passed to the pre-trained detection heads (at full resolution) with a batch size of 1. The detected images along with the detection outputs (in XML format) are stored for further accuracy calculation. Finally, the average precision per category (six categories) as well as the mean average precision, number of True Positives (TP), False Positives (FP), False Negatives (FNs) were computed using the procedure mentioned in [27], [54]

IV. RESULTS

In this section, we present detailed as well as summary results of the evaluation (see section III for details). First, Table 1, shows the detailed detection accuracy (average precision) results for each of the five categories present in test dataset i.e. Bottle, Car, Chair, Dining Table, Person and Potted Plant across the three detectors and seven HDR to LDR mapping techniques. Additionally, Table 1 also contains the per-object-category average precision (AP) results for each of the three detection models folded across seven HDR to LDR mapping techniques.

Table 1 not only presents a broad overview of the detection quality obtained from each of the three detectors, but also demonstrates the suitability of each of the seven HDR to LDR mapping techniques for object detection purposes under challenging lighting conditions.

Although Table 1 provides a detailed indication of the performance of various mapping techniques for object detection purposes, it does not provide a holistic and conclusive view of the evaluation results. Therefore, in Figure 7, Figure 8, and Figure 9 a summary of the detection results is presented folded across seven HDR to LDR mapping techniques and traditional HDR and three detection models.

Figure 7 shows a comparative heatmap of mAP values obtained for the seven HDR to LDR mapping techniques and traditional LDR across the three detection models used in this evaluation. This allows us to directly compare and contrast the suitability of each of the three detectors as well as the seven HDR mapping techniques in detecting objects under extreme lighting conditions, regardless of object category. In contrast, Figure 8 demonstrates the average precision results for each of the seven mapping techniques, folded across three detectors. This allows us to study the suitability of each mapping technique regardless of the detection model. Finally, for the purposes of detection, apart from image quality, Figure 9 looks at the overall inference time for

TABLE 1. Average precision per object category for each of the three object detectors and mean average precision folded across seven HDR to LDR mapping techniques and traditional LDR.

Object Detector	Mapping Technique	Object Category						
		Bottle	Car	Chair	Dinning Table	Person	Potted Plant	mAP
Faster RCNN	Ashikhmin	0.1	0.66	1.0	0.83	0.69	0.44	0.62
	Exposure Fusion	0.03	0.69	1.0	1.0	0.60	0.13	0.58
	Fattal	0.02	0.67	1.0	0.71	0.62	0.23	0.54
	Ferwerda	0.00	0.77	0.00	0.35	0.60	0.02	0.29
	iCAM06	0.03	0.69	1.00	1.00	0.60	0.13	0.57
	Mantiuk	0.07	0.72	1.00	0.49	0.70	0.36	0.56
	Reinhard	0.07	0.70	1.00	0.00	0.61	0.48	0.48
	LDR	0.08	0.78	1.00	0.28	0.60	0.11	0.47
	Average	0.05	0.70	0.86	0.63	0.63	0.26	-
R-FCN	Ashikhmin	0.10	0.70	1.00	0.00	0.70	0.14	0.44
	Exposure Fusion	0.00	0.67	1.00	0.06	0.58	0.00	0.38
	Fattal	0.01	0.67	1.00	0.01	0.63	0.00	0.39
	Ferwerda	0.00	0.78	0.00	0.00	0.55	0.04	0.23
	iCAM06	0.02	0.71	1.00	0.00	0.71	0.04	0.31
	Mantiuk	0.02	0.76	1.00	0.00	0.67	0.03	0.41
	Reinhard	0.07	0.79	1.00	0.00	0.63	0.02	0.42
	LDR	0.07	0.75	1.00	0.01	0.59	0.03	0.41
	Average	0.03	0.73	0.86	0.01	0.64	0.04	-
SSD	Ashikhmin	0.00	0.69	1.00	1.00	0.63	0.22	0.59
	Exposure Fusion	0.00	0.73	1.00	1.00	0.60	0.18	0.58
	Fattal	0.00	0.70	1.00	0.99	0.63	0.01	0.56
	Ferwerda	0.00	0.68	0.00	0.04	0.39	0.00	0.18
	iCAM06	0.00	0.78	1.00	1.00	0.64	0.03	0.57
	Mantiuk	0.00	0.76	1.00	0.94	0.62	0.17	0.58
	Reinhard	0.00	0.76	1.00	0.39	0.52	0.24	0.48
	LDR	0.00	0.77	0.53	0.95	0.54	0.05	0.47
	Average	0.00	0.73	0.86	0.77	0.58	0.12	-

each of the seven mapping techniques (folded across the three detection models), which helps us to determine the speed of each mapping technique, regardless of the detection model. However, the results demonstrated in Figure 9 are only indicative, as neither the mapping techniques nor detection models were optimised for the purposes of this study.

V. DISCUSSION

From Table 1, it becomes evidently clear that some of the overall suitability of the HDR to LDR mapping techniques are similar for detection purposes. This is corroborated by the results demonstrated in Figure 8 where we see similar accuracy across all mapping techniques apart from *FerwerdaTMO*. Also, from Figure 7, it can be observed that the HDR to LDR mapping techniques significantly outperform native LDR images in terms of detection accuracy. Hence, we can conclude that for detection purposes, the local and scene reproduction operators such as *Ashikhmin*, *Mantiuk* and *iCAM06* outperform others. The detection results are, however, more interesting. Overall, both Faster RCNN and SSD512 outperform R-FCN even though they have the same VGG 16 [39] feature extraction backbone. As expected, all detection models, regardless of the chosen mapping technique, are fairly accurate in detecting larger objects such as *car, chair* and *dining table* with accuracy dropping sharply in case on smaller objects such as *bottle* and *potted Plant*. Interestingly, the two-stage RPN based models, such as Faster RCNN and R-FCN, are marginally more accurate than SSD in detecting smaller objects. However, an anomaly can be seen in Table 1 where the category *dining table* is completely missed by R-FCN, even though both Faster RCNN and SSD 512 are fairly accurate in this category. Considering the similarity of Faster RCNN and R-FCN, this suboptimal result for this category might be attributed to default anchor box sizes and aspect ratios and might require further optimization for better results.

Furthermore, the paired results Figures 8 and 9, exhibit that even though *Ferwerda* has the lowest accuracy, the inference time is the least, thereby providing high FPS performance at the cost of image quality. Amongst the mapping techniques, *Ashikhmin* delivers the best feature quality for the detection models, thus facilitating high detection accuracy. This comes

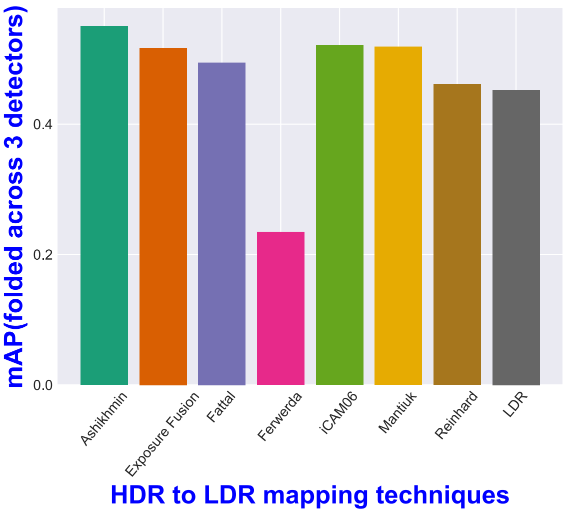


FIGURE 8. Mean Average Precision (mAP) results (folded across three detection models).

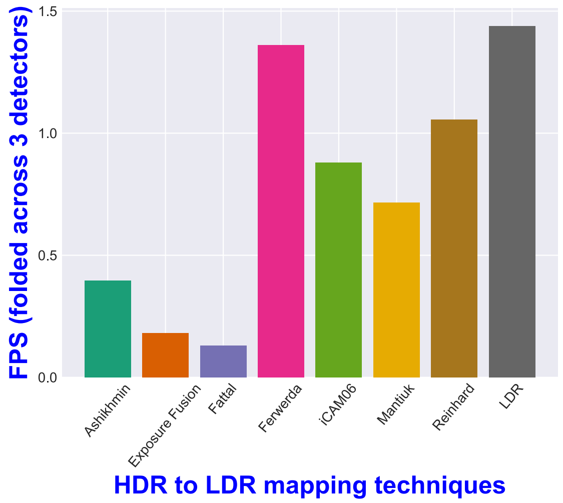


FIGURE 9. Inference speed in frames/sec (folded across three detection models).

at the cost of inference time. Since the eventual goal is find a balance between ''high accuracy'' and ''low inference time (high FPS)'', the paired results from Figures 8 and 9, seem to suggest that photographic tone mapping operator (*Reinhard* - global version) provides the best balance between both parameters. Finally, the heatmap in Figure 7 demonstrates that apart from *Ferwerda*, all mapping techniques typically outperform the traditional middle (0^{th}) exposure in terms of detection accuracy.

VI. LIMITATIONS

Although this manuscript presents an exhaustive feasibility study of several HDR to LDR mapping techniques for the purpose of object detection in challenging lighting conditions, there are a few limitations of this work.

First, in light of the recent proposals in object detection, the detection heads used in this work have been superseded by anchor-free detection heads such as CornerNet [4] and Objects as Points [5] both of which eliminate the need for anchor box design and optimization and thus could possibly result in better detection accuracy on the OOD dataset.

Second, the FPS figures obtained in this evaluation were overall turnaround per-frame which also includes I/O time. Also, since most of the HDR mapping techniques are implemented in an interpreted programming language i.e. MATLAB, the rest of the pipeline was also implemented in MATLAB to create a complete pipeline. A downside of this strategy is that neither the mapping techniques nor the detection framework are hardware/software optimised. Therefore, the results shown in Figure 9 are indicative and it can be expected that an optimised version of a reasonably fast and relatively straightforward mapping technique, such as *Reinhard*, when paired with an optimised detection model preferably implemented using mainstream deep learning frameworks [55], [56] would provide a more accurate indication of the real-world inference times.

Third, the settings for the HDR to LDR mapping techniques were in accordance to the previous studies in this area. However, none of the prior research was for detection purposes. As such, there might be better fine-tuned parameters which could result in a better detection accuracy.

Finally, due to the lack of native HDR footage, the size of the OOD dataset is a significant limiting factor for a comprehensive evaluation.

VII. CONCLUSION

The primary goal of this work was to study the feasibility of using existing HDR to LDR mapping techniques such that current detection models can be used to detect relevant objects under challenging lighting conditions. To that end, the most important contribution of this work is the detailed evaluation methodology presented in section III. In addition, we also conducted a comprehensive evaluation of the mapping techniques with three existing detection models to study the suitability and accuracy of existing detection models on tone-compressed HDR content (see section III-E for details). Results suggest that, although the performance of HDR to LDR mapping techniques are comparable, the local mapping and scene reproduction operators along with image appearance models tend to outperform other mapping techniques in detection quality. On the other hand, Faster RCNN and SSD 512 tend to outperform R-FCN with SSD being the fastest (least inference time) amongst the three. This is in accordance with the results presented in [35].

Comparing the HDR to LDR mapping techniques with the traditional middle expose, one can see that some of them outperform the middle exposure significantly. These results were obtained without any optimization of the mapping techniques for image detection. The results further suggest that optimization of mapping techniques for image detection could achieve even better results. Moreover, it is possible to build new mapping techniques that can enhance images characteristics and make them more suitable for object detection.

We believe that this work is one of the first steps towards object detection using HDR content. An interesting future work would be to train and use CNN based detection models directly on HDR content. Also, the creation of a larger HDR dataset is required before any further efforts to train detection heads with HDR content can be meaningfully undertaken.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

REFERENCES

[1] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," *Comput. Vis. Image Understand.*, vol. 117, no. 8, pp. 827–891, Aug. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731421300091X

[2] 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.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[4] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[5] X. Zhou, D. Wang, and P. Krähenbähl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[6] A. Artusi, F. Banterle, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2011.

[7] *Parameter values for the HDTV Standards for Production and International Programme Exchange*, document Recommendation itu-r bt.709, 2002. [Online]. Available: http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-5-200204-I!!PDF-E.pd

[8] *Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange*, document Recommendation itu-r bt.2020, 2015. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2020-2-201510-I!!PDF-E.pdf

[9] Y. Ou, P. Ambalathankandy, M. Ikebe, S. Takamaeda, M. Motomura, and T. Asai, "Real-time tone mapping: A state of the art report," 2020, *arXiv:2003.03074*. [Online]. Available: https://arxiv.org/abs/2003.03074

[10] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002, doi: 10.1145/566654.566575.

[11] A. Adams, "The negative: Exposure and development basic photo 2," *Morgan Lester*, vol. 98, p. 48, Oct. 1948.

[12] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," *ACM Trans. Graph.*, vol. 27, no. 3, p. 68, Aug. 2008. [Online]. Available: http://0-doi.acm.org.pugwash.lib.warwick.ac.uk/10.1145/1360612.1360667

[13] J. E. J. Tumblin and H. E. Rushmeier, "Tone reproduction for realistic computer generated images," Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep., 1991.

[14] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg, "Time-dependent visual adaptation for fast realistic image display," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 47–54.

[15] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 249–256, doi: 10.1145/566570.566573.

[16] W. T. Vetterling, S. A. Teukolsky, W. H. Press, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[17] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual adaptation for realistic image synthesis," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 249–258.

[18] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," *IEEE Comput. Graph. Appl.*, vol. 13, no. 6, pp. 42–48, Nov. 1993.

[19] G. Ward, "A contrast-based scalefactor for luminance display," *Graph. Gems*, vol. 4, pp. 415–421, Aug. 1994.

[20] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *Proc. 15th Pacific Conf. Comput. Graph. Appl.*, Oct. 2007, pp. 382–390.

[21] J. Kuang, G. M. Johnson, and M. D. Fairchild, "ICAM06: A refined image appearance model for HDR image rendering," *J. Vis. Commun. Image Represent.*, vol. 18, no. 5, pp. 406–414, Oct. 2007.

[22] M. D. Fairchild and G. M. Johnson, "Meet ICAM: A next-generation color appearance model," in *Proc. Color Imag. Conf.*, vol. 1, 2002, pp. 33–38.

[23] G. W. Meyer and D. P. Greenberg, "Perceptual color spaces for computer graphics," *ACM SIGGRAPH Comput. Graph.*, vol. 14, no. 3, pp. 254–261, Jul. 1980, doi: 10.1145/965105.807502.

[24] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 257–266, doi: 10.1145/566570.566574.

[25] R. W. Hunt, "An improved predictor of colourfulness in a model of colour vision," *Color Res. Appl.*, vol. 19, no. 1, pp. 23–26, 1994.

[26] E. Reinhard, E. A. Khan, A. O. Akyuz, and G. Johnson, *Color Imaging: Fundamentals Application*. Boca Raton, FL, USA: CRC Press, 2008.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[30] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[32] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*. [Online]. Available: http://arxiv.org/abs/1612.08242

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

[34] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055

[35] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[36] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[37] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: In defense of two-stage object detector," 2017, *arXiv:1711.07264*. [Online]. Available: http://arxiv.org/abs/1711.07264

[38] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, "ThunderNet: Towards real-time generic object detection on mobile devices," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6718–6727.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[43] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[44] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*. [Online]. Available: http://arxiv.org/abs/1805.04687

[45] *DSLR: Canon 5D Mark*. Accessed: Apr. 10, 2020. [Online]. Available: https://www.canon.pt/for_home/product_finder/cameras/digital_slr/eos_5d_mark_iii/

[46] *Spheron HDR Video Camera*. Accessed: Apr. 10, 2020. [Online]. Available: http://www.hdrv.org/HDRv.php

[47] A. Chalmers, G. Bonnet, F. Banterle, P. Dubla, K. Debattista, A. Artusi, and C. Moir, "High-dynamic-range video solution," in *Proc. ACM SIGGRAPH ASIA*, 2009, p. 71.

[48] ARRI. *Arri Alexa LF Video Camera*. Accessed: Apr. 11, 2020. [Online]. Available: https://www.arri.com/en/camera-systems/cameras/alexa-lf

[49] M. Ashikhmin, "A tone mapping algorithm for high contrast images," in *Proc. 13th Eurographics Workshop Rendering*, Cham, Switzerland, Switzerland, 2002, pp. 145–156. [Online]. Available: http://dl.acm.org/citation.cfm?id=581896.581916

[50] H. Zhao, X. Jin, and J. Shen, "Real-time tone mapping for high-resolution HDR images," in *Proc. Int. Conf. Cyberworlds*, Sep. 2008, pp. 256–262.

[51] A. Oäuz Akyáz, M. Levent Eksert, and M. Selin Aydin, "An evaluation of image reproduction algorithms for high contrast scenes on large and small screen display devices," *Comput. Graph.*, vol. 37, no. 7, pp. 885–895, Nov. 2013.

[52] M. Melo, M. Bessa, K. Debattista, and A. Chalmers, "Evaluation of tone-mapping operators for HDR video under different ambient luminance levels," *Comput. Graph. Forum*, vol. 34, no. 8, pp. 38–49, 2015.

[53] M. æadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Comput. Graph.*, vol. 32, no. 3, pp. 330–349, Jun. 2008.

[54] J. Cartucho. *Mean Average Precision*. Accessed: Apr. 13, 2020. [Online]. Available: https://github.com/Cartucho/mAP

[55] M. Abadi. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: http://tensorflow.org/

[56] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

RATNAJIT MUKHERJEE received the Ph.D. degree in high dynamic range video processing from the University of Warwick, in 2017. He was a Research Fellow with INESC TEC, Portugal. He currently works as an AI Researcher with the Advanced Research Laboratory, Navinfo Europe B.V., The Netherlands. His research interests include HDR video compression, computer vision, and artificial intelligence.

MIGUEL MELO is currently an Assistant Researcher in computer graphics with INESC TEC. He is the Manager of the Multisensory Virtual Reality Laboratory. His research interests include computer graphics, HDR, and multisensory virtual reality. He is also a Member of the Executive Committee of Eurographics.

VÍTOR FILIPE received the M.S. degree in informatics from the University of Minho, Portugal, in 1997, and the Ph.D. degree in electrical engineering from the University of Trás-os-Montes e Alto Douro, Portugal, in 2003. Since 2015, he has also been a Senior Researcher with INESC TEC. He is currently an Associate Professor with Habilitation in Electrical Engineering with the School of Science and Technology, UTAD. His research interests include computer vision, machine learning, and gait analysis.

ALAN CHALMERS received the M.Sc. degree (Hons.) from Rhodes University, in 1985, and the Ph.D. degree from the University of Bristol, in 1991. He is currently a Professor of visualisation with the WMG, University of Warwick, U.K., and a former Royal Society Industrial Fellow. He is an Honorary President of Afrigraph and a former Vice President of ACM SIGGRAPH. He has published over 250 papers in journals and international conferences on HDR, high-fidelity virtual environments, multi-sensory perception, parallel processing, and virtual archaeology. He successfully supervised 48 Ph.D. students. He is a U.K. Representative on IST/37 considering standards within MPEG.

MAXIMINO BESSA has been a Senior Researcher with INESC TEC and the Director of the Multisensory Virtual Reality Laboratory Massive, since 2009. He is currently an Assistant Professor with Habilitation with the Department of Engineering, University of Trás-os-Montes and Alto Douro, Portugal. Since 2003, he has been a Member of the Eurographics Association. From 2016 to 2020, he was the Vice-President of the Portuguese Computer Graphics Chapter.