Neural Auto-Exposure for High-Dynamic Range Object Detection

Emmanuel Onzon
Algolux

Fahim Mannan
Algolux

Felix Heide
Princeton University, Algolux

Abstract

Real-world scenes have a dynamic range of up to 280 dB that todays imaging sensors cannot directly capture. Existing live vision pipelines tackle this fundamental challenge by relying on high dynamic range (HDR) sensors that try to recover HDR images from multiple captures with different exposures. While HDR sensors substantially increase the dynamic range, they are not without disadvantages, including severe artifacts for dynamic scenes, reduced fill-factor, lower resolution, and high sensor cost. At the same time, traditional auto-exposure methods for low-dynamic range sensors have advanced as proprietary methods relying on image statistics separated from downstream vision algorithms. In this work, we revisit auto-exposure control as an alternative to HDR sensors. We propose a neural network for exposure selection that is trained jointly, end-to-end with an object detector and an image signal processing (ISP) pipeline. To this end, we use an HDR dataset for automotive object detection and an HDR training procedure. We validate that the proposed neural auto-exposure control, which is tailored to object detection, outperforms conventional auto-exposure methods by more than 6 points in mean average precision (mAP).

1. Introduction

From no ambient illumination at night to bright sunny day conditions, the range of possible luminances computer vision systems have to measure and analyze can exceed 280 dB, expressed here as the ratio of the highest to the lowest luminance value [54]. While the luminance range found at the same time in a typical outdoor scene is 120 dB, it is the “edge cases” that are challenging. For example, exiting a tunnel can include scene regions with almost no ambient illumination, the sun, and scene points with intermediate luminances, all in one image. Capturing this large dynamic range has been an open challenge for image sensing, and today’s conventional CMOS image sensors are capable of acquiring around 60-70 dB in a single capture [51]. This sensing constraint poses a fundamental problem for low-level and high-level vision tasks in uncontrolled scenarios, and it is critical for applications that base decision-making on vision modules in-the-wild, including outdoor robotics, drones, self-driving vehicles, driver assistance systems, navigation, and remote sensing.

To overcome this limitation, existing vision pipelines rely on HDR sensors that acquire multiple captures with different exposures of the same scene. A large body of work explores different HDR sensor designs and acquisition strategies [59, 8, 51], with sequential capture methods [67, 69, 47, 64] and sensors that split each pixel into two sub-pixels [66, 29, 30, 2] as the most successfully deployed HDR sensor architectures. Although modern HDR sensors are capable of capturing up to 140 dB at increasing resolutions, e.g., OnSemi AR0820AT, the employed multi-capture acquisition approach comes with fundamental limitations. As exposures are different in length or start at different times, dynamic scenes cause motion artefacts that are an open problem to eliminate [11, 63, 2]. Custom sensor architectures come at the cost of reduced fill-factor, and hence resolution, and sensor cost, compared to conventional intensity sensors. Moreover, capturing HDR images does not only require a sensor that can measure the scene but also necessitates optics for HDR acquisition, without glare and lens flare. Note also that in contrast to LDR sensors, interleaved HDR sensors cannot implement global shutter acquisition.

In this work, we revisit low dynamic range (LDR) sensors, paired with learned exposure control, as a computational alternative to the popular direction of HDR sensors. Existing auto-exposure (AE) control methods have been largely designed as proprietary compute blocks, often embedded by the sensor manufacturer on the same silicon as the sensor, producing perceptually pleasing images for human consumption. Conventional AE methods rely on image statistics [56, 4, 58], such as histogram or gradient statistics, and, as such, do not receive feedback from the task-specific vision module that ingests the camera images. Similarly, the vision module responsible for a higher-level vision task is designed, trained, and evaluated offline, often using JPEG images without any dependence on the live imaging pipeline [16, 43, 7]. We explore whether departing from conventional AE methods developed in isolation and instead learning a *neural exposure control that is* *jointly learned with a downstream vision module*, allows us to overcome the limitations of conventional LDR sensors and recent HDR sensors.

We propose a neural auto-exposure network that predicts optimal exposure values for a downstream object detection task. This control network and the downstream detector are trained in an end-to-end fashion jointly with a differentiable image processing pipeline between both models, which maps the RAW sensor measurements to RGB images ingested by the object detector model. The training of this end-to-end model is challenging as AE dynamically modifies the RAW sensor measurement. Instead of an online training approach which would require a camera and annotation in-the-loop, we train the proposed model by simulating the image formation model of a low-dynamic range sensor from input HDR captures. To this end, we acquire an automotive HDR dataset. We validate the proposed method in simulation and using an experimental vehicle prototype that evaluates detection scores for fully independent camera systems with different AE methods placed side-by-side and separately annotated ground truth labels. The proposed method outperforms conventional auto-exposure methods by 6.6 mAP points across diverse automotive scenarios.

Specifically, we make the following contributions:

- We introduce a novel neural network architecture that predicts exposure values driven by an object detection downstream network in real time.
- We propose a synthetic training procedure for our auto-exposure network that relies on a synthetic LDR image formation model.
- We validate the proposed method in simulation and on an experimental prototype, and demonstrate that the proposed neural auto-exposure control method outperforms conventional auto-exposure methods for automotive object detection across all tested scenarios.

2. Related Work

High Dynamic Range Imaging As existing sensors are not capable of capturing the entire range of luminance values in real-world scenes in a single shot, HDR imaging methods employ multiplexing strategies to recover this dynamic range from multiple measurements with different exposures [45, 9, 54]. These approaches can be combined with smart metering strategies ([15, 28, 36, 17]). For static scenes, conventional HDR acquisition methods rely on temporal multiplexing by sequentially capturing LDR images for different exposures and then combining them through exposure bracketing [45, 9, 54, 21, 49, 24]. These methods suffer from motion artefacts for dynamic scenes, which a large body of existing work addressed in post-capture stitching [36, 38, 14, 20, 27, 34, 57], optical flow [44], and deep learning [32, 33]. While these methods are successful for photography, they are, unfortunately, not real-time and leave high-resolution HDR imaging for robotics an open challenge. For safety-critical applications, including autonomous driving, recent work that hallucinates HDR content from LDR images [12, 13, 40, 41, 46] is not an alternative for detection and navigation stacks that must measure the real world. At the same time, HDR image processing pipelines have been manually designed and optimized in the past, in isolation from the downstream detector task [5]. Our work bridges this gap and optimizes camera control for challenging HDR scenarios, driven by a downstream task loss.

Adaptive Camera Control Although auto-exposure control is fundamental to acquisition with all conventional low-dynamic range sensors, especially when employed in dynamic outdoor environments, existing exposure control software (and auto-white balance control) has been largely limited to proprietary algorithms [53, 68]. This is because the feedback of exposure control algorithms must exceed real-time capture rates, and, as a results, exposure control algorithms are often implemented in hardware on the sensor or as part of the hardware image signal processing pipeline (ISP). Existing classical algorithms pose optimal exposure selection as an optimal control problem on image statistics [56, 4, 65], or they rely on efficient heuristics [39, 52]. A further successful direction solves model-predictive control problems [52, 60, 61] to predict optimal exposure values. Recently, a number of works select exposure values to optimize local image gradients [58, 10] instead of global image statistics. Although commodity smartphone devices rely heavily on semantic auto-exposure, especially driving portrait photography, only very few semantic auto-exposure methods are documented [71, 37]. Only recently, Yang et al. [70] have proposed to personalize semantic exposure control using reinforcement learning, performing similar to modern smartphone exposure control methods. Instead of tailoring exposure control to user preference, we address automotive exposure control that is optimized for a downstream perception task, such as object detection, driven by an end-to-end IoU loss.

Post-Capture Tonemapping A large body of work has explored tonal adjustments to high-dynamic range or low-dynamic range images after the capture process, driven by scene semantics [31, 6, 35, 48, 22]. Recent tonemapping approaches rely on deep convolutional neural networks [18, 42] to perform tonal and exposure-adjustments post-capture. While these approaches can compress dynamic range after captures, allowing to balance local gradient magnitudes that can be minute compared to the global intensity range of HDR images [54], they cannot recover details lost during the capture process, including saturated

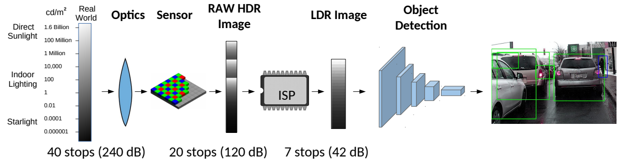

Figure 1: Dynamic range in the image processing pipeline.

and low-light flux-limited regions. In this work, we jointly train the exposure control with the post-processing and downstream network, allowing us to outperform existing automotive auto-exposure methods in all tested conditions for synthetic and diverse experimental driving campaigns. Recently, several works address parameter optimization [62, 50] for non-differentiable hardware ISPs and are orthogonal to the proposed method as they *exclude camera control*. In contrast, we learn auto-exposure control and rely on differentiable ISPs

3. Single-Shot Image Formation

Direct sunlight has a luminance around $1.6 \cdot 10^9\ \mathrm{cd/m^2}$, while starlight lies around $10^{-4}\ \mathrm{cd/m^2}$. Accordingly, the total range of luminances the human eye is exposed to ranges from $10^{-6}\ \mathrm{cd/m^2}$ to $10^8\ \mathrm{cd/m^2}$ which is a range of 280 dB. However, the range of differences that the eye can discern is lower, at 60 dB in very bright conditions (contrast ratio of 1000) and 120 dB in dimmer conditions (contrast ratio of 10^6), see [11] Chapter 15. The dynamic range of a camera employing a 12-bit sensor is bounded from above by 84 dB because of the bounded and quantized sensing, and we note that the effective dynamic range is even lower because of optical and sensor noises (around 60-70 dB) [54]. Examples of such optical noise sources are veiling glare, stray light and aperture ghosts. The sensor noise tends to dominate the optical noise for LDR cameras while the converse is true for HDR cameras. The dynamic range is progressively shrunk throughout the image processing pipeline (Figure 1). It follows that choosing where this dynamic range lies in the scale of possible luminances is critical to capture the useful information for the task at hand. This is the role of the AE.

The image formation model considered in this work is illustrated in Figure 2. Specifically, we consider the recording of a digital value by the sensor at a pixel as the result of the following single-shot capture process. Radiant power ϕ exposes the photosite during the exposure time t, creating $y_p(\phi \cdot t)$ photoelectrons. We express ϕ in electrons (e-) (following [25]) and t in seconds (s). Dark current creates $y_d(\mu_d)$ electrons, where μ_d is the average number of electrons in the absence of light. This measurement results in y_e accumulated electrons, that is

$$y_e = \max(y_p(\phi \cdot t) + y_d(\mu_d), M_{\text{well}}), \tag{1}$$

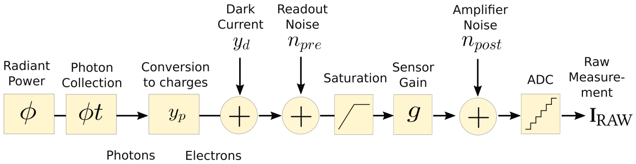

Figure 2: The irradiance at a photosite flows through a sequence of linear and nonlinear operations that result in a digital value which is the sensor RAW measurement. Each of these steps add noise and affects the overall measurement image quality.

where M_{well} is the full well capacity expressed in electrons.

These y_e electrons are converted to a voltage which is amplified before being converted to a digital number that is recorded by the sensor as a pixel value. The voltage is affected by noise before amplification (readout noise) and after amplification (analog-to-digital conversion noise).

This process results in the following model for raw pixel measurement, see also [25] and [3]. A value recorded by the sensor is expressed in digital numbers (DN), a dimensionless unit.

$$I_{\text{sensor}} = q(g \cdot (y_e + n_{\text{pre}}) + n_{\text{post}}), \tag{2}$$

where n_{pre} is the thermal and quantum noise introduced before amplification, and n_{post} is the readout noise introduced after and during amplification. Both n_{pre} and n_{post} are expressed in DN. The constant g is the camera gain and expressed in digital numbers per electron (DN/e-). It can be further broken down into $g = K \cdot g_1$, where g_1 is the gain at ISO 100 and K is the camera setting of the gain *e.g.*, $K = 1$ for ISO 100, $K = 2$ for ISO 200, bounded by the maximum analog gain. The function q is quantization performed by the analog-to-digital converter,

$$q(x) = \min\left(\lfloor x + 0.5 \rfloor, M_{\text{white}}\right), \tag{3}$$

where M_{white} is the white level *i.e.*, the maximum value that can be recorded by the sensor. Here we assume that the image of the target camera is recorded as a 12-bit raw image so we use $M_{\text{white}} = 2^{12} - 1$. For the purpose of training with stochastic gradient descent optimization we override the gradient of $\lfloor \cdot \rfloor$ (the floor function) as the function uniformly equal to 1 *i.e.*, the gradient is computed as if $\lfloor \cdot \rfloor$ was replaced by the identity function.

The model presented above differs from [25] and [3] in that the quantization is modeled explicitly with function q, while [25] and [3] model it as a quantization noise, which is included in the post-amplification noise n_{post}. However, we still express the quantization error as a variance when considering the signal-to-noise ratio (SNR). For a detailed derivation of the different noise quantities, SNR and dynamic range, see the Supplementary Material.

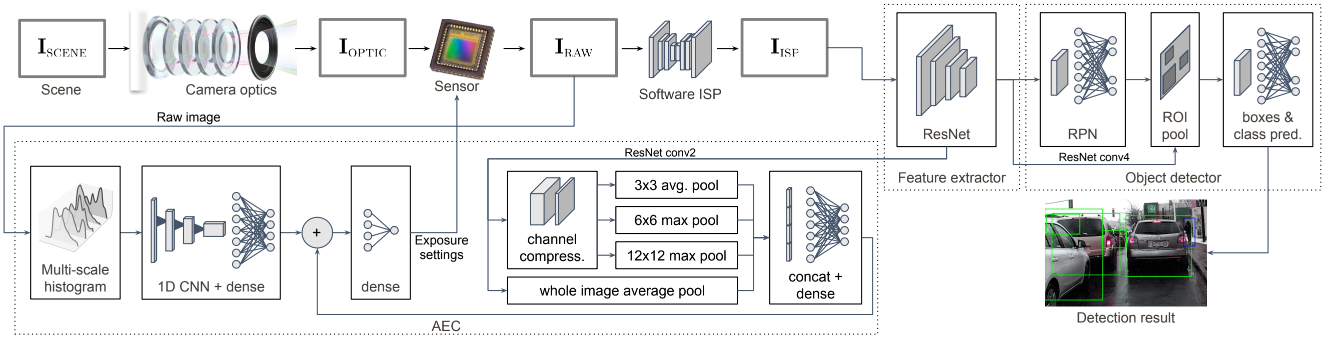

Figure 3: Overview of our end-to-end live object detection method with neural auto exposure control. The global image feature branch is shown on the left of the AE block and the semantic feature branch is illustrated on the right. In the pooling operations, n x n does not refer to a receptive field but means the feature map is divided up into a n by n array.

4. Learning Exposure Control

As a computational alternative to the popular direction of HDR sensors, in this section, we propose to revisit low dynamic range sensors, paired with learned exposure control. An illustration of the proposed method is shown in Figure 3. Specifically, given frame number t, the proposed learned exposure control network predicts the exposure and gain values of the next frame $(t + 1)$ from global image statistics and scene semantics in two network branches. The first branch, "Histogram NN", operates on a set of histograms computed from the image at three different scales. While this branch efficiently encodes global image features, the second branch "Semantic NN" exploits semantic features that are shared with a downstream object detector module. Both global and semantic features are summed together to form a joint feature vector and a head predicts the final exposure value from it. The two branches can either be used indepnedently or jointly. We refer to the joint model as "Hybrid NN". In the following, we describe the two network branches.

4.1. Global Image Feature Branch

To incorporate global image statistics without the need for a network with a very large receptive field, we rely on histogram statistics as input to the first branch of the proposed learned auto-exposure method. Specifically, this branch takes as input a set of histograms at 3 different scales. We note that histogram statistics can be estimated with efficient ASIC blocks on the sensor or in a co-processor [1]. At the intermediate scale, and respectively the finest scale, the image is divided into a 3 by 3 and a 7 by 7 array of sub images. At the coarsest scale we consider the whole image. From each of these 59 images a 256-bin histogram is computed based on the first green pixel of the Bayer pattern. These histograms are stacked together to form an input to the neural network histogram branch with shape [256, 59]. To predict auto-exposure values, we use a six-layer neural network. The first three layers are 1D convolutional layers with kernel size 4 and stride 4. The last three layers are fully connected. A full architecture definition can be found in the Supplementary Material.

Each of the layers 1 to 5 are followed by a ReLU activation function. The last layer is followed by a custom activation function that computes the final exposure adjustment for frame t as:

$$u_t = \exp(2 \cdot (\text{sigmoid}(x) - 0.5) \cdot \log(M_{\text{exp}})) \qquad (4)$$

where x is the preactivation of layer 6. The constant $M_{\text{exp}} > 0$ is the maximum exposure change, it is a bound such that $u_t \in [M_{\text{exp}}^{-1}, M_{\text{exp}}]$. In our implementation we set $M_{\text{exp}} = 10$.

4.2. Semantic Feature Branch

The second branch of the proposed method incorporates semantic feedback into the auto-exposure control. To this end, we reuse the computation of the feature extractor of the object detector from the current frame. We use the output of ResNet conv2 (see [26]) as the input to our semantic feature branch. We first apply channel compression from 128 to 26 channels and refer to the output as the compressed feature map (CFM). Then we apply pyramid pooling at 4 scales. At the coarsest of the four scales we apply average pooling of the output of conv2 along the two spatial dimensions. At the finest scales we use a growing size of max and average pooling operations on the CFM. A full architecture definition can be found in the Supplementary Material. We flatten and concatenate the tensors of the previous pooling operations, which is followed by a densely connected layer. The two output feature vectors, from each of the branches, are summed followed by a common densely connected layer with a custom activation function as described in Section 4.1.

4.3. Exposure Prediction Filtering

To further improve the accuracy of the exposure control at inference time, we aggregate exposure predictions across consecutive frames with an exponential moving average of

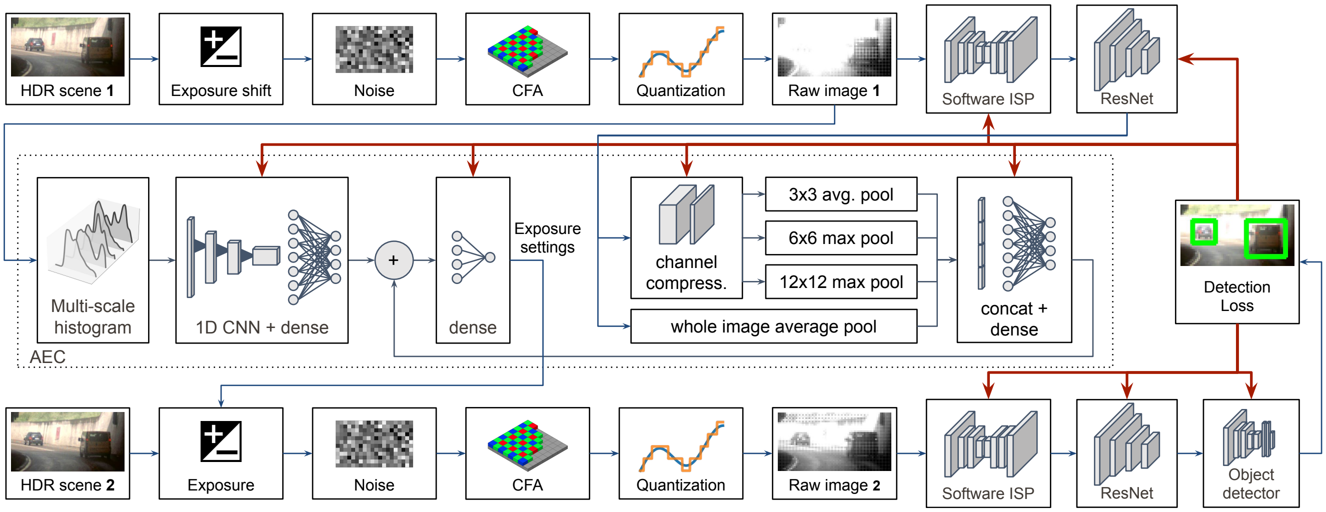

Figure 4: Overview of our end-to-end training methodology with learned AE and object detection. The red arrows indicate the trainable parameter and weight updates. The two instances of ResNet and the ISP instances share their weights.

the logarithm of the exposure,

$$\log e_t = \mu \cdot \log e_{t-1} + (1-\mu) \cdot \log\left(e_{t-1} \cdot u_t\right) \tag{5}$$

i.e. with $e_t = e_{t-1} \cdot u_t^{1-\mu}$, e_t is the next exposure value, e_{t-1} is the exposure for the previous frame, u_t is the exposure adjustment predicted by the neural networks of Sections 4.1 and 4.2. We set the smoothing hyperparameter to $\mu = 0.9$ in our implementation.

4.4. Shutter Speed and Gain from Exposure Value

The neural exposure prediction described above produces a single exposure value $e_t = K \cdot t_{\text{exp}}$ with the gain K and the exposure time t_{exp}. Since maximizing the exposure time maximizes the SNR (see Supplemental Material), it is

$$K = \max(1, e_t / T_{\max}), \quad t_{\text{exp}} = e_t / K \tag{6}$$

where $T_{\max}$ is the maximum allowed exposure time, which we set to $T_{\max} = 15\text{ms}$.

5. Training

An overview of our training approach is illustrated in Figure 4. In the following, we describe the training methodology in detail.

5.1. HDR Training Dataset

The proposed training pipeline simulates LDR raw images from HDR captures. The HDR image data takes the form of 3 LDR JPEG images that are combined at training time to form a linear color image. JPEG images are convenient to save disk space and time when loading training examples, rather than using the 24 bit linear images directly to make the dataset. The training dataset consists of 1600 pairs of images that have been acquired using a test vehicle and using a camera with a Sony IMX490 HDR image sensor. Each pair of images consists of two successive frames of which the second one has been manually annotated for automotive 2D object detection. We refer to the Supplemental Material for additional details. About 50% of the images have been taken during day time, 20% at dusk and 30% at night time, with diverse weather conditions. The driving locations include urban and suburban areas, countryside roads and highways. The raw HDR data was processed by a state-of-the-art ARM Mali C71 ISP to obtain three LDR images. These images are rescaled to the resolution of the target image sensor (Sony IMX249) and saved in the sRGB color space.

5.2. LDR Image Capture Simulation

The proposed AE model is trained on LDR raw images simulated using the image formation model from Sec. 3. Specifically, we calibrate the sensor noise parameters and set the camera gain, K, and exposure time t, see Supplemental Material for details.

The irradiance ϕ for each pixel of the image is simulated using images taken by the HDR camera described above. This is done by taking the three JPEG encoded LDR images whose combined dynamic range covers the full 140 dB of the HDR image. More specifically, for each LDR image $\mathbf{J}_i$, the scaled linear image is, $\mathbf{I}_i = \alpha_i \cdot \varphi(\mathbf{J}_i)$. Here the exposure factor $\alpha_i = (K_i \cdot t_i)^{-1}$ is decreasing with i, and φ is the inverse tonemapping operator to recover a linear image in $[0, 1]$. Hence, each image $\mathbf{I}_i$ has values in the range $[0, \alpha_i]$.

Latent HDR Image. A linear HDR image $\mathbf{I}_{\text{hdr}}$ is produced from the n scaled linear light images by computing the minimum variance unbiased estimator (following [25]), *i.e.*, the weighted average of pixel values across the set of LDR images with weights equal to the inverse of the noise variance,

$$\mathbf{I}_{\text{hdr}} = \frac{\sum_{i=1}^{3} w_i \cdot \mathbf{I}_i}{\sum_{i=1}^{3} w_i} \quad \text{with} \quad w_i = \frac{\delta_{I_i < M_{\text{white}}}}{\alpha_i^2 \cdot V_{\text{unsat}}}, \tag{7}$$

where V_{unsat} is the variance of unsaturated pixels, see Supplemental Material for a derivation.

Irradiance Simulation. We simulate the irradiance per pixel ϕ_{sim} with the help of the linear light HDR image $\mathbf{I}_{\text{hdr}}$ described above, that is $\phi_{\text{sim}} := \text{Bayer}(\gamma \cdot \mathbf{I}_{\text{hdr}})$. Here, Bayer is the Bayer pattern sampling of the image sensor. The conversion factor γ maps DN to the corresponding irradiance.

Noise simulation. Sensor noise is simulated at training time to match the distribution of the target LDR sensor. Since the captured data already contains noise, we add only the amount that reproduces the target sensor's noise characteristic through *noise adaptation*. We also apply *noise augmentation* for each training example by randomly varying the strength of the simulated noise around the noise strength targeted by noise adaptation, see Supplemental Material.

5.3. Network Training

During training, a single example is made up of two consecutive frames forming a mini sequence along with bounding boxes and class labels for the second frame, see Figure 4.

The full training pipeline consists of the following six steps. We first simulate a 12-bit capture for the first frame with a random exposure e_{rand} shifted from a base exposure e_{base} by a shift factor κ_{shift}, that is $e_{\text{rand}} = \kappa_{\text{shift}} \cdot e_{\text{base}}$. The base exposure e_{base} is computed adaptively from the HDR frame pixel values as $e_{\text{base}} = 0.5 \cdot M_{\text{white}} \cdot (\gamma \cdot \bar{\mathbf{I}}_{\text{hdr}})^{-1}$, with $\bar{\mathbf{I}}_{\text{hdr}}$ as the mean value of $\mathbf{I}_{\text{hdr}}$. The logarithm of κ_{shift} is sampled uniformly in $[\log 0.1, \log 10]$.

We then predict an exposure change with the proposed network using the given frame as input, and we simulate a 12-bit capture of the next frame with this adjusted exposure. The resulting frame is then processed by an ISP and an object detector predicts bounding boxes on the output RGB image of the ISP. The entire imaging and detection pipeline is supervised only with the object detector loss at the end. We use the same loss as in Girshick et al. [19, 55], but also add a weighted penalty on the L2 norm of the weights of the AE neural network. All steps are implemented with TensorFlow graphs such that the auto exposure network can be trained based on the object detector loss. For brevity, we refer the reader to the Supplemental Material for additional training details.

We note that, even with histograms alone, the other components of the pipeline (ISP, feature extractor, object detector) are *trained jointly* with the AE model, such that no optimal exposure exists for a given training example *i.e.*, the training cannot be done without the object detector and the computer vision task loss in the loop.

Table 1: Object detection performance (AP at IoU 0.5) for three exposure shift simulation scenarios, for 6 classes and mean AP accross classes (mAP). The base exposure is shifted by a factor randomly sampled in {0.667, 1.5} for small shifts, {0.25, 4} for moderate shifts and {0.1, 10} for large shifts. Results within one standard deviation of the corresponding best result are highlighted with *.

Method	**Classes**						**mAP**
	Bike	Bus & Truck	Car & Van	Person	Traffic Light	Traffic Sign	
GRADIENT AE [58]	17.56	31.26	60.70*	28.92	21.90	30.07	31.73
AVERAGE AE	16.01	29.74	59.56	28.85	21.53	29.70	30.90
HISTOGRAM NN *(ours)*	19.87*	33.11	60.43	29.55	22.60	**31.42**	32.83
SEMANTIC NN *(ours)*	**20.19**	34.15	60.87*	30.21*	23.35	30.87	33.27
HYBRID NN *(ours)*	20.18*	**37.06**	**61.07**	**30.60**	**23.98**	31.18*	**34.01**
			Mild exposure shift $k = 1.5$				
GRADIENT AE [58]	17.02	25.47	57.27	24.93	20.87	27.95	28.92
AVERAGE AE	15.50	29.09	58.08	27.17	21.29	28.63	29.96
HISTOGRAM NN *(ours)*	19.80	33.99	60.32	29.41	22.69	31.34*	32.92
SEMANTIC NN *(ours)*	19.76	32.55	60.72*	30.38*	23.50	**31.41**	33.05
HYBRID NN *(ours)*	**20.29**	**37.29**	**61.22**	**30.44**	**23.95**	31.28*	**34.08**
			Moderate exposure shift $k = 4$				
GRADIENT AE [58]	13.22	19.81	48.00	18.61	16.18	21.62	22.91
AVERAGE AE	12.99	25.10	53.83	23.81	18.62	26.30	26.77
HISTOGRAM NN *(ours)*	18.32	32.06	60.39	28.44	22.70	**31.12**	32.17
SEMANTIC NN *(ours)*	17.65	26.82	60.19	28.97	23.20	30.75	31.26
HYBRID NN *(ours)*	**19.42**	**35.18**	**61.01**	**29.81**	**23.70**	30.96*	**33.35**
			Large exposure shift $k = 10$				

Table 2: Impact of fine tuning in the training pipeline.

Method	$k = 1.5$	$k = 4$	$k = 10$
GRADIENT AE pretrained on LDR dataset	19.73	17.71	13.36
AVERAGE AE pretrained on LDR dataset	18.94	18.02	15.35
GRADIENT AE fine tuned on HDR dataset without AE	31.66	27.13	20.91
AVERAGE AE fine tuned on HDR dataset without AE	31.10	29.41	25.37
GRADIENT AE fine tuned on HDR dataset with AE	31.73	28.92	22.91
AVERAGE AE fine tuned on HDR dataset with AE	30.90	29.96	26.77

6. Evaluation

In this section, we assess the proposed learned auto-exposure method and compare it to existing baseline algorithms. Evaluating auto-exposure algorithms requires image acquisition with the predicted exposure, or a simulation of the capture process. To this end, we first validate the method on capture simulations in Table 1, allowing us to emulate the identical sensor irradiance present at the sensor. For the experimental comparisons in Table 3, we employ completely separate camera systems, each controlled by different free-running auto-exposure algorithms in real-time, and mount them side-by-side in a capture vehicle. The proposed method outperforms existing auto-exposure methods both in simulation and experimentally.

6.1. Synthetic Assessment

We first evaluate the proposed method by simulating scene intensity shifts using captured HDR data. To this end, we use a dataset of 400 pairs of consecutive HDR frames taken with the HDR Sony IMX490 sensor that was also used for capturing the training set. We apply noise adaptation,

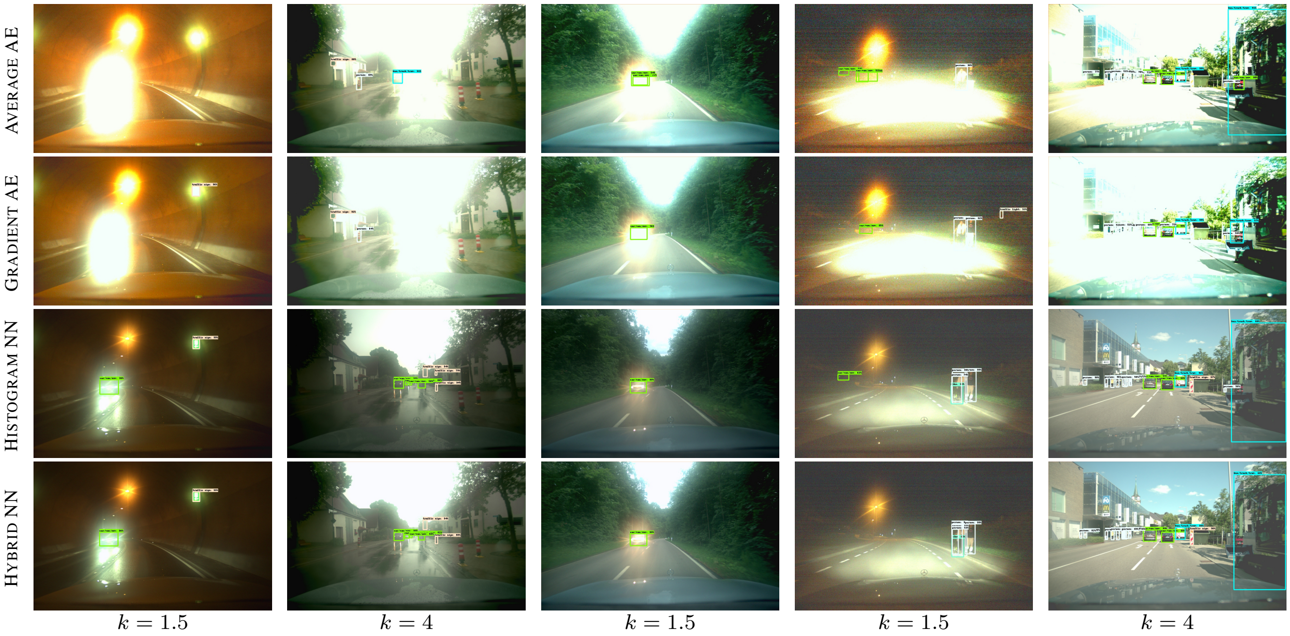

Figure 5: Comparison of the two proposed methods and the two baselines, see text, using simulations of mild ($k = 1.5$) and moderate ($k = 4$) exposure shifts.

but no noise augmentation, see Supplemental Material. For each pair of frames, we simulate a random test exposure in the same way as in the training pipeline except here κ_{shift} is sampled with equal probabilities in the set $\{k^{-1}, k\}$, with $k = 1.5$ for mild shifts, $k = 4$ for moderate shifts and $k = 10$ for large shifts. The evaluation metric is the object detection average precision (AP) at 50% IoU over the 400 pairs and their horizontal flip. For each tested AE method and each $k \in \{1.5, 4, 10\}$, the experiments are repeated 12 times and we compute the mean and the standard deviation of the AP score. For fair comparisons, we fine-tune the detector networks separately for all auto-exposure baselines.

Quantitative and Qualitative Validation. We compare five AE algorithms. The three proposed algorithms *i.e.*, each of the two branches proposed as standalone and the hybrid model, are compared along with two baseline algorithms, an average-based AE algorithm [1], and an AE algorithm [58] driven by local image gradients. The average-based AE employs an efficient and fast scheme [1] that adjusts the mean pixel value I_{mean} of the current raw frame and adjusts the exposure by a factor $0.5 \cdot M_{\text{white}}/I_{\text{mean}}$. The gradient-based AE from Shim et al. [58] aims to adjust exposure to maximize local image gradients. We use the proposed parameters $\lambda = 1000$, $\delta = 0.06$, and $K_p = 0.5$. Both baseline algorithms (see Supplemental Document) are implemented using TensorRT and run in real-time on a Nvidia GTX 1070.

Table 1 lists the average precision (AP) of all compared algorithms for six automotive classes, see evaluation details in the Supplemental Document. These synthetic results validate the proposed method as it outperforms the two baseline algorithms for each of the 6 classes and across all three exposure shift scenarios, with a larger margin for larger shifts. While the semantic branch outperforms the global image feature branch for smaller shifts, the opposite is true for larger shifts. The hybrid model takes advantage of the complementarity of both branches and outperforms the single branch models for all exposure shifts. Figure 5 shows qualitative comparisons confirming that the proposed method can recover from extreme exposures in cases where the baselines fail.

Table 2 compares the baseline algorithms when finetuning on the HDR dataset with and without the AE in the pipeline and when pretraining on the LDR dataset only. These results show that even non-trainable AE algorithms can benefit from the proposed training pipeline.

6.2. Experimental Assessment

We validate the proposed method experimentally by implementing the proposed method and best baseline AE algorithm from the simulation section on two separate camera prototype systems that are mounted side-by-side in a test-vehicle. The captured frames from the same automotive scenes but different camera systems are annotated manually and separately for a fair comparison.

Prototype Vehicle Setup. We compare the object detection results of the proposed method (hybrid model) with the average AE baseline method. Each of the two cameras is free-running and takes input image streams from separate imagers mounted side-by-side on the windshield of a vehicle, see Figure 7. Images are recorded with the object detector and each AE algorithm running live. For fair com-

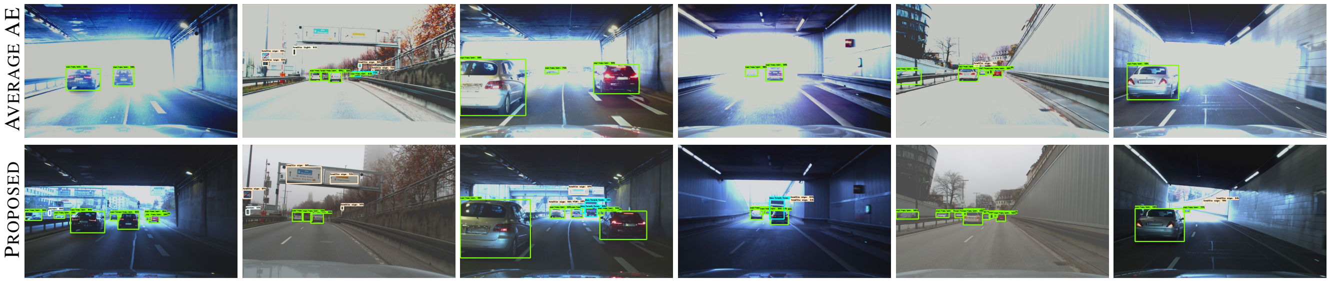

Figure 6: Experimental prototype results of the proposed neural AE compared to the Average AE baseline method, see text, using the real-time side-by-side prototype vehicle capture system shown in Figure 7. The proposed method accurately balances exposure of objects still in the tunnel with exposure of objects outside of the tunnel and adapts robustly to changing conditions.

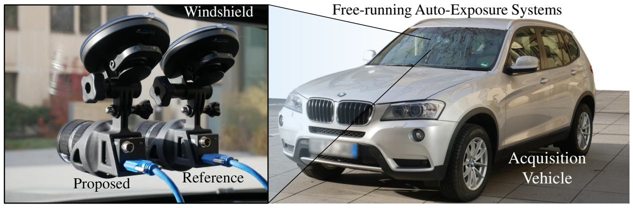

Figure 7: Side-by-side capture setup for the experimental comparison of the proposed Hybrid NN with the average-based auto-exposure control baseline, see text.

parisons, we use the individually finetuned detector with the average AE baseline method. All compared AE methods and inference pipelines run in real-time on two separate machines, each equipped with a Nvidia GTX 1070 GPU.

The driving scenarios are highway and urban ones in European cities during the daytime. We include several tunnels in the test set to also assess conditions of rapidly changing illumination. The route is taken two times during two successive days at the same time of the day. The input to the pair of compared algorithms is *swapped between the two drives*, such that the algorithm receiving input from the left camera the first day receives input from the right camera the second day and conversely. A total of 3140 frames is selected for testing each AE algorithm. Frames are selected in pairs, one from each algorithm, such that they match the sampling time. The selected test frames are annotated for four of the six classes listed in Section 6.1.

Quantitative and Qualitative Validation. All separately acquired images were manually annotated by humans for the automotive classes that the models were trained for. Using these ground-truth annotations, the detection performance of each pipeline is evaluated as shown in Table 3. These results confirm the improvement in object detection using the proposed model in both simulation and real-world experiments.

Table 3: Experimental object detection evaluation for the proposed hybrid NN and the average-based AE method running side-by-side in the prototype vehicle from Figure 7. The reported scores are the average precision at IoU 0.5 for each of the 4 classes and the mean across classes.

Method	Classes				mAP
	Bike	Bus & Truck	Car & Van	Person	
AVERAGE AE	11.93	28.92	54.20	20.17	28.80
HYBRID NN *(ours)*	**13.96**	**34.09**	**58.90**	**22.53**	**32.37**

Figure 6 shows a qualitative comparison that further validate the proposed method in challenging high dynamic range conditions. Specifically, the method is capable of carefully balancing the exposure between dark and bright objects even in rapidly changing conditions.

For additional comparisons to HDR exposure selection and fusion, we implemented the method from Gupta et al. [23] and compare to it in the Supplemental Document.

7. Conclusion

Exposure control is critical for computer vision tasks as under or overexposure can lead to significant image degradations and signal loss. Existing HDR sensors and reconstruction pipelines approach this problem by aiming to acquire the full dynamic range of a scene with multiple captures of different exposures. This brute-force capture approach has the downside that these captures are challenging to merge for dynamic objects, and sensor architectures suffer from reduced fill-factor. In this work, we revisit low dynamic range (LDR) sensors, paired with learned exposure control, as a computational alternative to the popular direction of HDR sensors. Existing auto-exposure control methods have been largely restricted to proprietary ASIC blocks, prohibiting access to the vision community. This work proposes a neural exposure control method that is optimized for downstream vision tasks and makes use of the scene semantics to predict optimal exposure parameters. We validate the effectiveness of our approach in simulation and experimentally in a prototype vehicle system, where the proposed neural auto-exposure outperforms conventional methods by more than 6 points in mean average precision. In the future, we envision joint optimization of the sensor architecture itself along with the proposed exposure control as an exciting step towards learning the cameras of tomorrow.

References

[1] *ARM Mali C71*, 2020 (accessed Nov 11, 2020). 4, 7

[2] T Asatsuma, Y Sakano, S Iida, M Takami, I Yoshiba, N Ohba, H Mizuno, T Oka, K Yamaguchi, A Suzuki, et al. Sub-pixel architecture of cmos image sensor achieving over 120 db dynamic range with less motion artifact characteristics. In *Proceedings of the 2019 International Image Sensor Workshop*, 2019. 1

[3] European Machine Vision Association. Emva standard 1288, standard for characterization of image sensors and cameras, release 3.1. 2016. 3

[4] Sebastiano Battiato, Arcangelo Ranieri Bruna, Giuseppe Messina, and Giovanni Puglisi. *Image processing for embedded devices*. Bentham Science Publishers, 2010. 1, 2

[5] Michael S Brown and SJ Kim. Understanding the in-camera image processing pipeline for computer vision. 2015. 2

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[8] Arnaud Darmont. *High dynamic range imaging: sensors and architectures, second edition*. 2019. 1

[9] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '08*, 1997. 2

[10] Zhushun Ding, Xin Chen, Zhe Jiang, and Cheng Tan. Adaptive exposure control for image-based visual-servo systems using local gradient information. *JOSA A*, 37(1):56–62, 2020. 2

[11] Frédéric Dufaux, Patrick Le Callet, Rafal Mantiuk, and Marta Mrak. *High dynamic range video: from acquisition, to display and applications*. Academic Press, 2016. 1, 3

[12] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017. 2

[13] Konstantina Fotiadou, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Snapshot high dynamic range imaging via sparse representations and feature learning. *IEEE Transactions on Multimedia*, 2019. 2

[14] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–7, 2009. 2

[15] Orazio Gallo, Marius Tico, Roberto Manduchi, Natasha Gelfand, and Kari Pulli. Metering for exposure stacks. In *Computer Graphics Forum*, volume 31, pages 479–488. Wiley Online Library, 2012. 2

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[17] Natasha Gelfand, Andrew Adams, Sung Hee Park, and Kari Pulli. Multi-exposure imaging on mobile devices. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 823–826, 2010. 2

[18] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):118, 2017. 2

[19] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[20] Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free hdr reconstruction. *ACM Trans. Graph.*, 32:201:1–201:10, 2013. 2

[21] Michael D. Grossberg and Shree K. Nayar. High dynamic range from multiple images: Which exposures to combine? 2003. 2

[22] Dong Guo, Yuan Cheng, Shaojie Zhuo, and Terence Sim. Correcting over-exposure in photographs. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 515–521. IEEE, 2010. 2

[23] Mohit Gupta, Daisuke Iso, and Shree K Nayar. Fibonacci exposure bracketing for high dynamic range imaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1473–1480, 2013. 8

[24] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2010. 2

[25] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560. IEEE, 2010. 3, 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[27] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1170, 2013. 2

[28] Kun-Fang Huang and Jui-Chiu Chiang. Intelligent exposure determination for high quality hdr image generation. In *2013 IEEE International Conference on Image Processing*, pages 3201–3205. IEEE, 2013. 2

[29] S Iida, Y Sakano, T Asatsuma, M Takami, I Yoshiba, N Ohba, H Mizuno, T Oka, K Yamaguchi, A Suzuki, et al. A 0.68 e-rms random-noise 121db dynamic-range sub-pixel architecture cmos image sensor with led flicker mitigation. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 10–2. IEEE, 2018. 1

[30] Manuel Innocent, Angel Rodriguez, Deb Guruaribam, Muhammad Rahman, Marc Sulfridge, Swarnal Borthakur,

Bob Gravelle, Takayuki Goto, Nathan Dougherty, Bill Desjardin, et al. Pixel with nested photo diodes and 120 db single exposure dynamic range. In *International Image Sensor Workshop*, pages 95–98, 2019. 1

[31] Neel Joshi, Wojciech Matusik, Edward H Adelson, and David J Kriegman. Personal photo enhancement using example images. *ACM Trans. Graph.*, 29(2):12–1, 2010. 2

[32] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36:144:1–144:12, 2017. 2

[33] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. *Comput. Graph. Forum*, 38:193–205, 2019. 2

[34] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B. Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Trans. Graph.*, 32:202:1–202:8, 2013. 2

[35] Sing Bing Kang, Ashish Kapoor, and Dani Lischinski. Personalization of image enhancement. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1799–1806. IEEE, 2010. 2

[36] Sing Bing Kang, Matthew Uyttendaele, Simon A. J. Winder, and Richard Szeliski. High dynamic range video. *ACM Trans. Graph.*, 22:319–325, 2003. 2

[37] Wen-Chung Kao, Chien-Chih Hsu, Chih-Chung Kao, and Shou-Hung Chen. Adaptive exposure control and real-time image fusion for surveillance systems. In *2006 IEEE international symposium on circuits and systems*, pages 4–pp. IEEE, 2006. 2

[38] Erum Arif Khan, Ahmet Oguz Akyüz, and Erik Reinhard. Ghost removal in high dynamic range images. *2006 International Conference on Image Processing*, pages 2005–2008, 2006. 2

[39] June-Sok Lee, You-Young Jung, Byung-Soo Kim, and Sung-Jea Ko. An advanced video camera system with robust af, ae, and awb control. *IEEE Transactions on Consumer Electronics*, 47(3):694–699, 2001. 2

[40] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018. 2

[41] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[42] Tzu-Mao Li, Michaël Gharbi, Andrew Adams, Frédo Durand, and Jonathan Ragan-Kelley. Differentiable programming for image processing and deep learning in halide. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 2

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[44] Ce Liu. Exploring new representations and applications for motion analysis. 2009. 2

[45] Steve Mann and Rosalind W. Picard. Being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures. 1994. 2

[46] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *CoRR*, abs/1803.02266, 2018. 2

[47] Mitsuhito Mase, Shoji Kawahito, Masaaki Sasaki, Yasuo Wakamori, and Masanori Furuta. A wide dynamic range cmos image sensor with multiple exposure-time signal outputs and 12-bit column-parallel cyclic a/d converters. *IEEE Journal of Solid-State Circuits*, 40(12):2787–2795, 2005. 1

[48] Belen Masia and Diego Gutierrez. Content-aware reverse tone mapping. In *2016 International Conference on Artificial Intelligence: Technologies and Applications*. Atlantis Press, 2016. 2

[49] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Comput. Graph. Forum*, 28:161–171, 2009. 2

[50] Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, and Felix Heide. Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7529–7538, 2020. 3

[51] Jun Ohta. *Smart CMOS image sensors and applications*. CRC press, 2020. 1

[52] SangHyun Park, GyuWon Kim, and JaeWook Jeon. The method of auto exposure control for low-end digital camera. In *2009 11th International Conference on Advanced Communication Technology*, volume 3, pages 1712–1714. IEEE, 2009. 2

[53] Jonathan B. Phillips and Henrik Eliasson. *Camera Image Quality Benchmarking*. Wiley Publishing, 1st edition, 2018. 2

[54] Erik Reinhard, Greg Ward, Summant Pattanaik, Paul E. Debevec, Wolfgang Heidrich, and Karol Myszkowski. High dynamic range imaging: Acquisition, display, and image-based lighting. 2010. 1, 2, 3

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 6

[56] Simon Schulz, Marcus Grimm, and Rolf-Rainer Grigat. Using brightness histogram to perform optimum auto exposure. *WSEAS Transactions on Systems and Control*, 2(2):93, 2007. 1, 2

[57] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31:203:1–203:11, 2012. 2

[58] Inwook Shim, Tae-Hyun Oh, Joon-Young Lee, Jinwook Choi, Dong-Geol Choi, and In So Kweon. Gradient-based camera exposure control for outdoor mobile platforms. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6):1569–1583, 2018. 1, 2, 6, 7

[59] Arthur Spivak, Alexander Belenky, Alexander Fish, and Orly Yadid-Pecht. Wide-dynamic-range cmos image sensors - comparative performance analysis. *IEEE transactions on electron devices*, 56(11):2446–2461, 2009. 1

[60] Yuanhang Su and C-C Jay Kuo. Fast and robust camera's auto exposure control using convex or concave model. In *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pages 13–14. IEEE, 2015. 2

[61] Yuanhang Su, Joe Yuchieh Lin, and C-C Jay Kuo. A model-based approach to camera's auto exposure control. *Journal of Visual Communication and Image Representation*, 36:122–129, 2016. 2

[62] Ethan Tseng, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.*, 38(4):27–1, 2019. 3

[63] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: A survey and evaluation. In *Computer Graphics Forum*, volume 34, pages 683–707. Wiley Online Library, 2015. 1

[64] Sergey Velichko, Scott Johnson, Dan Pates, Chris Silsby, Cornelis Hoekstra, Ray Mentzer, and Jeff Beck. 140 db dynamic range sub-electron noise floor image sensor. *Proceedings of the IISW*, 2017. 1

[65] Quoc Kien Vuong, Se-Hwan Yun, and Suki Kim. A new auto exposure and auto white-balance algorithm to detect high dynamic range conditions using cmos technology. In *Proceedings of the world congress on engineering and computer science*, pages 22–24. San Francisco, USA: IEEE, 2008. 2

[66] Trygve Willassen, Johannes Solhusvik, Robert Johansson, Sohrab Yaghmai, Howard Rhodes, Sohei Manabe, Duli Mao, Zhiqiang Lin, Dajiang Yang, Orkun Cellek, et al. A 1280× 1080 4.2 μm split-diode pixel hdr sensor in 110 nm bsi cmos process. In *Proceedings of the International Image Sensor Workshop, Vaals, The Netherlands*, pages 8–11, 2015. 1

[67] Orly Yadid-Pecht and Eric R Fossum. Wide intrascene dynamic range cmos aps using dual sampling. *IEEE Transactions on Electron Devices*, 44(10):1721–1723, 1997. 1

[68] Lucie Yahiaoui, Jonathan Horgan, Senthil Yogamani, Ciaran Hughes, and Brian Deegan. Impact analysis and tuning strategies for camera image signal processing parameters in computer vision. In *Irish Machine Vision and Image Processing conference (IMVIP)*, 2011. 2

[69] David XD Yang and Abbas El Gamal. Comparative analysis of snr for image sensors with enhanced dynamic range. In *Sensors, cameras, and systems for scientific/industrial applications*, volume 3649, pages 197–211. International Society for Optics and Photonics, 1999. 1

[70] Huan Yang, Baoyuan Wang, Noranart Vesdapunt, Minyi Guo, and Sing Bing Kang. Personalized exposure control using adaptive metering and reinforcement learning. *IEEE transactions on visualization and computer graphics*, 25(10):2953–2968, 2018. 2

[71] Ming Yang, Ying Wu, James Crenshaw, Bruce Augustine, and Russell Mareachen. Face detection for automatic exposure control in handheld camera. In *Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, pages 17–17. IEEE, 2006. 2