Neural Exposure Fusion for High-Dynamic Range Object Detection

Emmanuel Onzon[1] Maximilian Bömer[1] Fahim Mannan[1] Felix Heide[1,2]
[1]Torc Robotics [2]Princeton University

Abstract

Computer vision in unconstrained outdoor scenarios must tackle challenging high dynamic range (HDR) scenes and rapidly changing illumination conditions. Existing methods address this problem with multi-capture HDR sensors and a hardware image signal processor (ISP) that produces a single fused image as input to a downstream neural network. The output of the HDR sensor is a set of low dynamic range (LDR) exposures, and the fusion in the ISP is performed in image space and typically optimized for human perception on a display. Preferring tonemapped content with smooth transition regions over detail (and noise) in the resulting image, this image fusion does typically not preserve all information from the LDR exposures that may be essential for downstream computer vision tasks. In this work, we depart from conventional HDR image fusion and propose a learned task-driven fusion in the feature domain. Instead of using a single companded image, we introduce a novel local cross-attention fusion mechanism that exploits semantic features from all exposures – learned in an end-to-end fashion with supervision from downstream detection losses. The proposed method outperforms all tested conventional HDR exposure fusion and auto-exposure methods in challenging automotive HDR scenarios.

1. Introduction

A wide range of computer vision tasks requires predictions in outdoor scenarios at real-time rates, with applications ranging from self-driving vehicles and advanced driver assistance systems to drones and robots in farming and outdoor maintenance. The global dynamic range of luminances in real-world scenes is 280 dB, see [34]. Inside this range, a typical outdoor scene covers a sub-range of about 120 dB, and such a typical scene already exceeds what conventional CMOS image sensors can capture at around 60-70 dB. And yet, in-the-wild computer vision systems, routinely have to handle more challenging conditions, like facing the sun in presence of large, shadow-casting objects (backlights) or moving from indoor to outdoor and back (*e.g.*, entrance and exit of a tunnel). In such cases, the range of luminances seen at the same time can reach 180 dB, and they exceed the range of today's robotic and automotive high dynamic range (HDR) image sensors (covering around 120-140 dB). Moreover, existing computer vision systems must also be able to adapt to changing illumination conditions in real-time, for example, when the vision system, or when large objects in the environment move quickly.

The traditional approach to tackle these challenges is to employ an HDR image sensor coupled with a hardware image signal processor (ISP) and an auto-exposure control system, each of them having been designed independently. More precisely, the HDR image sensor captures multiple exposures that are fused and processed in image space by an ISP. The output of the ISP is a single HDR color image which is consumed by a computer vision module that has been designed and trained independently of the other components in the pipeline. Each individual capture in this pipeline, acquired at a different exposure, covers a low dynamic range (LDR), *i.e.*, not exceeding 70 dB per image, while the total dynamic range covered by the set of LDR images covers a larger dynamic range. The fusion algorithm that produces the image output in image space (*i.e.*, the *fused image*), is typically designed in isolation from the other components of the vision pipeline. In particular, it is not optimized for the computer vision task at hand, be that detection, segmentation, or localization.

In this work, we depart from the conventional approach of capturing a bracketed HDR raw capture, fusion and detection. Instead of image-space HDR fusion, we propose a *feature-domain fusion approach driven by a downstream detection task*, without needing to reconstruct a single HDR image, see Figure 3. Specifically, we propose an early fusion approach that fuses feature maps from the different exposures into a single feature map. We introduce a novel attention module to help the neural network determine, at each spatial location, the exposure that contains the most relevant information concerning the object detection task. To this end, we devise a *local cross-attention fusion* block which attends to features locally across exposures. The queries of this cross-attention module are learned, while the keys and the values are the feature vectors at each location, across the different exposures.

The intuition behind our local cross-attention module is best conveyed by the attention maps it produces. Figure 2 shows an example of such learned attention maps paired with the corresponding images. The maps illustrate that

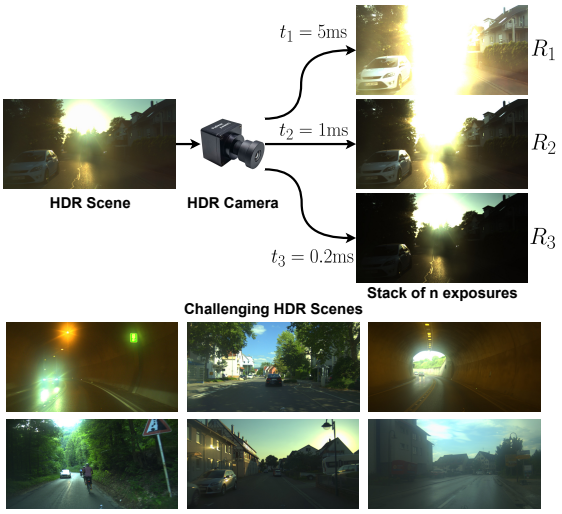


Figure 1. Modern HDR image sensors are capable of producing a stack of LDR images taken at different exposures in a short time frame. We take advantage of this feature in our method to perform fusion at a later stage in the pipeline. At the bottom of the figure, we show challenging scenarios for conventional HDR systems: Tunnel entrance and exit, oncoming traffic or strong backlight. Scenes with large luminance range complicate HDR fusion in image space and result in poor detail and low contrast.

our attention module locally gives more weight to features from semantically relevant and adequately exposed regions. In this particular example, we see that the lowest exposure feature map weights the area outside the tunnel higher than inside, while the medium and high exposures emphasize the inside regions of the tunnel. The weights are distributed evenly across exposures for areas with task-irrelevant semantic content, such as the sky, while in areas with high information content, the weights are concentrated in the best-exposed exposure.

We train our multi-exposure vision pipeline end-to-end, including a differentiable ISP and feature domain exposure fusion. This training is driven by detection losses typically used in object detection training pipelines [8, 35]. We learn feature-based fusion for $n > 1$ sub-exposure captures (see Figure 1) jointly with the ISP and the object detector, and we demonstrate that this *outperforms existing object detection approaches based on HDR fusion*. We validate the proposed feature-domain exposure fusion with a test set of automotive scenarios, and we compare the method against existing HDR reconstruction methods. The proposed method outperforms the conventional exposure fusion methods by 2.7% mAP. We validate all algorithm choices with extensive ablations experiments that test different feature-domain HDR fusion strategies. Specifically, we make the following contributions:

- We propose a novel neural feature-space fusion approach inside the detection model, as an alternative to *image space* exposure fusion for HDR object detection.

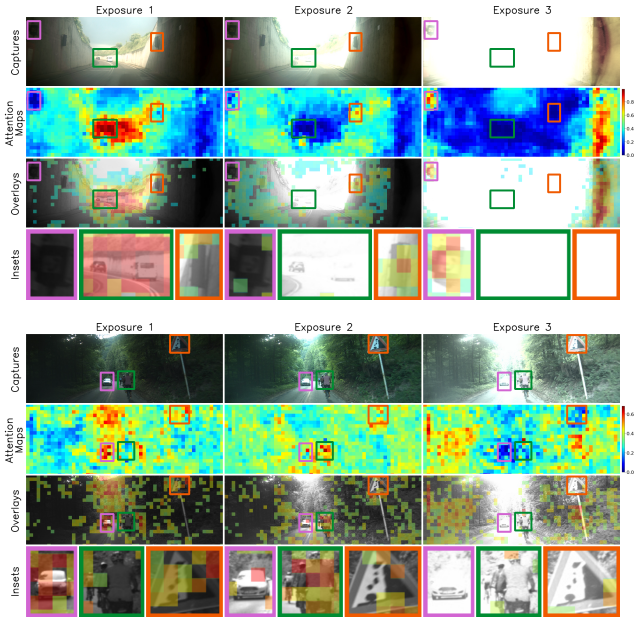


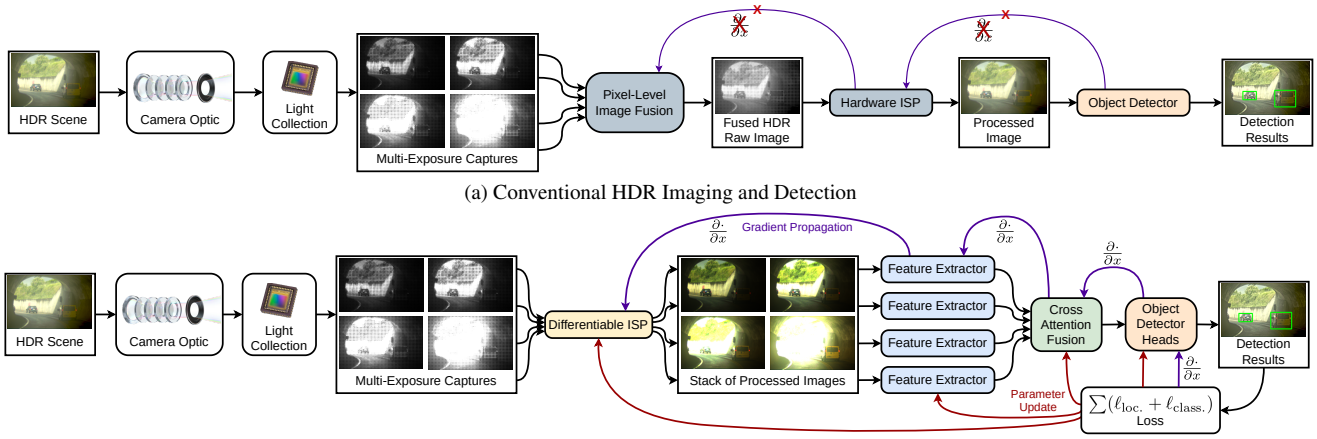
Figure 2. Multi-exposure captures and attention maps that illustrate the behavior of the proposed local-cross attention mechanism, see text. Only the highest weight per exposure is overlaid to highlight which exposure is given more weight at each spatial location.

- We design a new type of attention module, *local cross-attention fusion*, to perform feature fusion and is driven by a downstream detection task.
- We validate that the proposed method outperforms existing image space methods for automotive object detection across all tested scenarios and over all tested methods: +2.3% (*Deep HDR*) and +2.7% (*Raw HDR*).

2. Related Work

Next, we review related work on HDR imaging using exposure fusion, object detection, and learned HDR. Most prior works that we discuss primarily treat HDR imaging and perception as independent tasks which can lead to failure in high contrast scenes, see Fig. 1.

Exposure Fusion for High Dynamic Range Imaging The dynamic range of real-world scenes is much greater than what current sensors cover, and therefore a single exposure is insufficient for most real-world driving scenarios (e.g., tunnel entrances and exits). Exposure fusion is one of several strategies for capturing the large range of illuminations with multiple exposures [3, 25, 34]. Single exposure capture cameras typically apply image-dependent metering strategies to capture the largest dynamic range possible [6, 7, 15, 17], while multi-exposure cameras rely on temporal multiplexing of different exposures to obtain a single HDR image [3, 9, 11, 25, 28, 34]. Our work explores fusion in the feature domain, driven by a detection loss, with-



(a) Conventional HDR Imaging and Detection

(b) Proposed Neural Exposure Fusion for Detection

Figure 3. Conventional *HDR exposure fusion* is done in image space, before object detection. We propose an alternative approach to HDR object detection, where multi-exposure captures are not merged on the sensor but *fused in the feature domain*. The proposed pipeline reasons on features from separate exposures and relies on an attention module to fuse them together, which is trained end-to-end along with all the other modules of the pipeline, driven by the detection loss.

out needing to reconstruct a single HDR image.

Object Detection Networks Object detection networks can be classified into single-stage and two-stage meta-architectures based on how the inputs are chosen for object classification and regression [14]. In single-stage models [23, 24, 33], each cell of the feature map is considered for potential object category with different bounding box sizes then further refined and classified. During the first stage of a two-stage detector, the feature map is used for detecting regions of interest where objects can potentially be found. The potential regions are then cropped and fed to a detection head that does the final bounding box regression and classification [2, 22, 35]. We deliberately demonstrate the proposed method with the popular Faster-RCNN [35] meta-architecture with a custom lightweight 28-layer ResNet [13] backbone split into two stages.

Learning-based HDR Imaging and Perception Deep learning for HDR has primarily investigated generating HDR from a single LDR [4, 5, 18, 20, 26], HDR from multi-exposure fusion of LDR [16, 19, 32, 40, 41], and learned capture techniques [27, 29, 31]. A review of recent deep learning-based HDR method can be found in [39]. A few works propose to combine HDR imaging with perception tasks. For example, [38] proposes traffic light detection in dual-channel HDR image where the dark channel is used for detection and bright channel for classification, [30] proposed HDR object detection by converting HDR to LDR images. Some methods [1, 21] consider two differently exposed HDR stereo images for depth estimation. In contrast to these works, the proposed fusion of the individual sub-exposures is done in the feature domain, guided by a downstream loss, instead of the image domain as in conventional HDR fusion [34].

3. HDR Image Formation

In this section, we briefly review conventional image space exposure fusion. Typical HDR image pipelines produce an HDR raw image I_{HDR} by fusing n LDR images $R_1, \ldots, R_n$, that is

$$I_{\text{HDR}} = \text{ExpoFusion}(R_1, \ldots, R_n). \quad (1)$$

The LDR images $R_1, \ldots, R_n$ are recorded sequentially (or simultaneously using separate photo-sites per pixel) as n different recordings of the radiant scene power ϕ_{scene}. Specifically, an image R_j, $j \in \{1, \ldots, n\}$, with exposure time t_j and gain setting K_j is

$$R_j = \min\left((\phi_{\text{scene}} \cdot t_j + n_{\text{pre}}) \cdot g \cdot K_j + n_{\text{post}}, M_{\text{white}}\right), \quad (2)$$

where g is the conversion factor of the camera from radiant energy to digital number for unit-gain, n_{pre} and n_{post} are the pre and post-amplification noises, and M_{white} is the white level, *i.e.*, the maximum sensor value that can be recorded.

The fused HDR image is formed as a weighted average of the LDR images,

$$I_{\text{HDR}} = \sum_{j=1}^{n} w_j \cdot R_j, \quad (3)$$

where the w_j are per-pixel weights such that pixels that are saturated are given a zero weight. The role of the weights is to merge content from different regions of the dynamic range in a way that reduces artifacts, in particular noise. A popular approach is to choose the weights w_j such that I_{HDR} is the minimum variance unbiased estimator [12].

4. Neural Exposure Fusion

In this section, we first formalize conventional HDR pipelines (see Fig. 3a) before introducing the proposed

method (see Fig. 3b). In conventional HDR pipelines, the HDR image results from the fusion of n LDR raw images which are recorded in a burst following an exposure bracketing scheme. Such an image-space exposure fusion is designed *independently* of the vision task. As an alternative, we investigate here *feature-space exposure fusion* where we produce features from all exposures before fusing them based on semantic information. In addition, feature fusion is supervised by the downstream object detection loss. In other words, the training of the feature fusion module is driven by the performance on the object detection task. More formally, we can express a conventional HDR object detection pipeline (see Fig. 3a) as the following composition of operations

$$(b_i, c_i, s_i)_{i\in\mathcal{I}} = \mathrm{OD}(\mathrm{ISP}_{\mathrm{hw}}(\mathrm{ExpoFusion}(R_1, \ldots, R_n))), \tag{4}$$

where the b_i are the detected bounding boxes and c_i and s_i the corresponding inferred classes and confidence scores, the symbols OD, $\mathrm{ISP}_{\mathrm{hw}}$ and ExpoFusion denote the object detector, the hardware ISP and the image space exposure fusion, and $R_1, \ldots, R_n$ are the raw LDR images recorded by the HDR image sensor. Note that the exposure fusion outputs a *single* image that is ingested by the downstream pipeline to extract features. In contrast, we propose the following feature-space fusion

$$(b_i, c_i, s_i)_{i\in\mathcal{I}} = \mathrm{ODH}\Big(\mathrm{Fusion}\Big(\mathrm{FE}(\mathrm{ISP}(R_1)), \ldots, \mathrm{FE}(\mathrm{ISP}(R_n))\Big)\Big).$$

Here, we do not use a fused HDR image. Instead, we learn to extract *features for each exposure* that are fused in feature-space. The operator FE is the feature extraction, and ODH is the downstream part of the object detector, *i.e.*, the object detector heads. We share weights between the different feature extraction branches. The symbol Fusion denotes the neural fusion, which fuses feature maps from different exposures. We rely on a differentiable ISP in our method for each of the n raw subexposures $R_1, \ldots, R_n$. The entire model is trained *end-to-end as a differentiable multiexposure HDR capture and vision pipeline* where exposure control, ISP, feature extraction, fusion and the heads of the object detector are trained jointly. Specifically, for the object detector, we use the Faster-RCNN [35] meta-architecture with a 28-layer variant of ResNet [13] as feature extractor.

Note that, departing from the standard single-exposure method, we capture and extract features from *multiple different* exposures and propose a feature-based fusion for these multiple exposures. As such, the fusion of the information extracted from the different exposures is critical for the proposed model to be effective, which we discuss in the following.

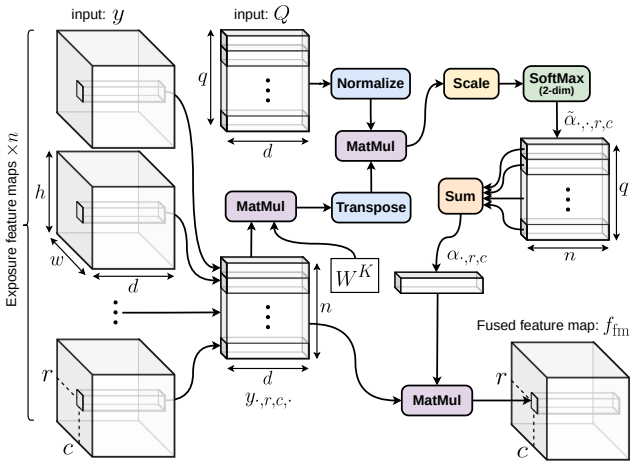


Figure 4. Local Cross-Attention Fusion. Cross attention with learned query matrix Q is applied locally to the n feature maps stacked into y, at row r and column c, resulting in the weight vector $\alpha_{\cdot,r,c}$, which is used to produce the vector at location (r, c) of the fused feature map f_{fm}. In contrast to [37], the softmax is normalized with respect to both axes.

4.1. Exposure Feature Fusion with Local Cross-Attention

We propose to fuse exposure features with local cross-attention, illustrated in Fig 4. To this end, we first extract a stack of feature maps for each exposure, that is, for exposure j, the following

$$y_{j,r,c,k} = \left(\mathrm{FE}(\mathrm{ISP}(R_j))\right)_{r,c,k}, \tag{5}$$

where r, c are spatial coordinates, and $k \in \{1, \ldots, d\}$ is a feature channel resulting from feature extractor FE.

We fuse the n exposure feature maps corresponding to the different exposures together by performing a locally weighted average with local weights that are computed according to attention maps. These attention maps are stacked together across the first dimension axis into the tensor $\alpha_{j,r,c}$. The fused feature map f_{fm} is then computed by applying a weighted sum reduction across the first dimension, that is

$$f_{\mathrm{fm},(r,c,k)} = \sum_{j=1}^{n} \alpha_{j,r,c} \cdot y_{j,r,c,k}. \tag{6}$$

To predict $\alpha_{j,r,c}$, we design a new attention module which we call *local cross-attention fusion*. In our attention module, the keys and the values are feature vectors from the feature maps y and the queries are trainable parameters. The module attends to the feature vectors across exposures and across a set of learned queries.

This means that for each row r and column c of the feature map, we consider the matrix $y_{\cdot,r,c,\cdot}$ such that for $1 \leqslant j \leqslant n$ and $1 \leqslant k \leqslant d$, $(y_{\cdot,r,c,\cdot})_{j,k} = y_{j,r,c,k}$, and apply our attention module locally to produce the n-dimensional local attention weight vector $\alpha_{\cdot,r,c}$,

$$\alpha_{\cdot,r,c} = \mathrm{Attention}(Q, y_{\cdot,r,c,\cdot}), \tag{7}$$

where Q is a learned query matrix of shape (q, d), which is shared across all locations (r, c).

Specifically, for each row $r = 1, \ldots, h$ and each column $c = 1, \ldots, w$ we first compute the key matrix $K^{(r,c)}$ at location (r, c) with the following matrix multiplication

$$K^{(r,c)} = y_{\cdot,r,c,\cdot} W^K, \tag{8}$$

where W^K is the learnable projection matrix with d rows and d columns for the keys (as in [37]). Queries are directly learned, the redundant projection step is therefore skipped.

As an intermediate step, we compute the *expanded attention map* $\tilde{\alpha}$ of shape (q, n, h, w). At each location (r, c) we represent α by the matrix $\tilde{\alpha}_{\cdot,\cdot,r,c}$, which is computed as

$$\tilde{\alpha}_{\cdot,\cdot,r,c} = \text{softmax}\left(\frac{\tilde{Q}(K^{(r,c)})^T}{\sqrt{d}}\right). \tag{9}$$

The matrix $\tilde{Q}$ is the normalized version of matrix Q, where each row of Q has been divided by its ℓ^2 norm. The result of the above softmax operation is a matrix of shape (q, n). We note that in this softmax operation, we apply the normalization of the softmax with respect to both axis of the matrix, *i.e.*, if z is a $q \times n$ matrix,

$$\text{softmax}(z)_{i,j} = \frac{e^{z_{i,j}}}{\sum_{i'=1}^{q}\sum_{j'=1}^{n} e^{z_{i',j'}}}, \tag{10}$$

while in the conventional query-key-value attention modules [37], the normalization is only applied with respect to the second axis.

Finally, the stacked attention map α is obtained by summing the expanded attention map $\tilde{\alpha}$ along its first axis, that is

$$\alpha_{j,r,c} = \sum_{i=1}^{q} \tilde{\alpha}_{i,j,r,c}. \tag{11}$$

Here, the final step to compute the fused feature map, *i.e.*, Equation (6), is analog to the matrix multiplication between the softmax output and the value matrix in [37]. Hence, we can view our cross-attention fusion as a variant of the query-key-value attention module, where the role of the value is played by $y_{\cdot,r,c,\cdot}$. While a projection matrix W^V is applied to the value matrix in [37], we do not apply any here.

4.2. Post Fusion Processing

The fused feature map can be used as input to any object detector that relies on a feature map. In our implementation, we adopt the corresponding first and second stages of [35]. That is to say, the fused feature map is input to the Region Proposal Network (RPN, see [35]), as well as to the ROI pooling operation (see [8]), and the remaining downstream processing closely follows [35] until producing the final bounding boxes and class scores.

In Section 6, we evaluate our method with different fusion variants, see Section 6.2 for a brief description and the Supplement for the details.

4.3. Differentiable ISP

The image signal processor (ISP) for all fusion strategies is composed of a sequence of conventional ISP modules, that permit differentiability, with the following processing steps: contrast stretching, demosaicing, image resizing, color correction, low frequency denoising, sharpening, contrast enhancing. We refer to the Supplemental Document for additional details. We implement all ISP blocks as differentiable operations to backpropagate through them, and we note that other differentiable ISP modules could be used.

4.4. Exposure Control

To set the exposure values – note that this is a problem separate from exposure fusion addressed in this work – we generalize the neural exposure control module from [31] to a set of n exposures. During inference, the module predicts exposure shifts to capture n LDR images. The model is trained using a synthetic training procedure. We provide details about the architecture and training procedure of the adapted module in the Supplemental Document.

5. Training

For training and testing of the proposed method, we use a dataset of automotive HDR images captured with the Sony IMX490 sensor mounted on a test vehicle. The sensor produces images that are 24 bits when decompanded. The training dataset consists of 18790 examples.

To train our end-to-end HDR object detection method, we simulate the capture of n 12-bit LDR raw images $R_1, \ldots, R_n$, from each HDR image I_{HDR} of the training set. These n raw images are used as input to our object detection pipeline. The object detection loss from [35] and its gradients are computed. The parameters of all trainable modules, including feature fusion, are updated with a gradient descent step. Following [31], exposure control is also factored into the training loop. See the Supplemental Material for details.

For our HDR baselines (see Section 6.1) instead of simulating n 12-bit raw images, we simulate a single 20-bit HDR image, reproducing the dynamic range of existing automotive HDR sensors like the OnSemi AR0231.

For our comparisons in the next section, we use the same starting point for training, *i.e.* a pretrained ISP – object detector pipeline. For additional details about the training methodology, we refer to the Supplemental Material.

6. Evaluation

In this section, we validate the proposed method. We compare different variants of the proposed fusion approach to conventional HDR Imaging and detection pipelines, and alternative fusion approaches, in diverse HDR scenarios.

Table 1. HDR object detection evaluation for different neural exposure fusion strategies compared to conventional LDR/HDR Imaging and object detection pipelines. Our proposed feature fusion strategies result in significant gains in mAP compared to pixel-level fusion methods, and improvements in most of the six considered object classes.

Method	Point of Fusion	Classes						mAP
		Bike	Bus & Truck	Car & Van	Person	Traffic Light	Traffic Sign	
Shim et al. [36] (LDR)	N/A	9.3	5.5	27.7	16.3	14.7	14.1	14.6
Onzon et al. [31] (LDR)	N/A	23.9	15.3	72.1	43.4	39.4	52.2	41.1
Raw HDR	Pre-ISP	23.4	15.2	72.1	43.7	41.8	52.8	41.5
Debevec and Malik [3]	Post-ISP	23.0	15.0	71.9	44.9	44.1	51.1	41.7
Deep HDR [16]	Post-ISP	25.6	**16.7**	72.2	44.6	43.4	48.7	41.9
Hanji et al. (PPNE) [10]	Pre-ISP	26.4	16.4	72.4	46.0	43.9	53.4	43.1
Max Pooling Fusion *(ours)*	Conv4	26.1	15.8	73.9	46.2	42.6	54.8	43.2
Conv 1 x 1 Fusion *(ours)*	Conv4	20.9	12.9	70.8	41.4	38.0	46.9	38.5
Conv 3 x 3 Fusion *(ours)*	Conv4	20.8	13.3	70.1	39.4	37.0	46.7	37.9
Late Fusion *(ours)*	2nd Stage	**27.5**	14.2	73.8	**47.2**	42.8	52.3	43.0
Local Cross Attention *(ours)*	Conv4	26.8	16.6	**74.3**	47.0	**44.4**	**56.3**	**44.2**

Our test set consists of 1996 pairs of consecutive HDR frames taken under a variety of challenging conditions. The second frame of each mini sequence is manually annotated with 2D bounding boxes. The examples are distributed across the following different illumination categories: sunny, cloud/rain, backlight, tunnel, dusk, night.

We simulate captures using 7 exposure shifts $\kappa_{\text{shift}} \in 2^{\{-15,-10,-5,0,5,10,15\}}$, that represent different evaluation scenarios of varying difficulty. The evaluation metric is the object detection average precision (AP) at 50% IoU, which is computed for the full test set. For fair comparison, all tested pipelines use the learned auto-exposure from [31], with minimal modifications to support 3-exposure inputs for the methods that require it (see Section 4.4).

6.1. Baseline Detection Pipelines

We compare against recent HDR and LDR baseline methods. Specifically, we compare against two image space exposure fusion methods: A custom HDR strategy *RAW HDR*, and *Deep HDR* [16]. Both pipelines synthesize an HDR Image by performing pixel fusion. Both variants convert the raw input image to 32 bit floating point and use the same differentiable ISP module (see 4.3) and object detector and are jointly finetuned on the training dataset for fair comparison. We compare to two LDR object detection pipelines that differ in their exposure selection approach. LDR Gradient AE uses the method from [36], and the second LDR object detection pipeline follows the method of [31], which is conceptually closest to our method. To our knowledge there are no existing feature-space exposure fusion approaches that are optimized for downstream task performance that we can include in our ablations. In the following, we describe the HDR pipelines that we compare to.

6.1.1 RAW HDR

For this baseline, no feature fusion is performed, instead a single 20-bit raw HDR image is simulated, as explained in Section 5. This image is then used as input to the ISP.

Table 2. Ablations experiments with varying number of exposures using Local Cross Attention (LCA) and experimental validation of its internal components.

Method	Point of Fusion	Classes						mAP
		Bike	Bus & Truck	Car & Van	Person	Traffic Light	Traffic Sign	
SoftMax 1D	Conv4	26.7	16.6	73.6	45.9	42.5	53.6	43.2
Skip Normalize Q	Conv4	25.7	15.8	71.4	45.3	41.8	53.8	42.3
Skip Multiply W^K	Conv4	26.5	17.2	74.0	46.9	44.2	55.9	44.1
LCA 2 Exposures	Conv4	27.1	16.5	74.1	46.7	42.3	55.7	43.7
LCA 4 Exposures	Conv4	26.7	16.7	74.4	47.9	44.3	56.2	44.4

6.1.2 Deep HDR

For this baseline, no feature fusion is performed, instead the LDR captures are processed independently by the ISP and merged to an HDR Image at the end of the ISP. The fusion is performed following the approach from [16].

6.2. Alternative Fusion Strategies

We validate the proposed method by comparing to the following alternative fusion strategies. We describe them briefly below, see Supplemental Material for details.

6.2.1 Maximum Pooling Fusion Strategy

For this strategy, the feature maps are fused together at the end of the feature extractor, as in Section 4.1. We employ a variant of the local cross-attention fusion, where we make a drop-in replacement of the local cross-attention fusion module by a maximum reduction across the n exposures, *i.e.*, we consider the following fused feature map,

$$f_{\max,(r,c,k)} = \max_{j=1,\dots,n} y_{j,r,c,k}. \tag{12}$$

6.2.2 Convolutional Fusion Strategy

For this strategy again, the feature maps are fused together at the end of the feature extractor, as in Section 4.1. The local cross attention fusion module is replaced by a convolutional layer. Specifically, the feature maps corresponding to the n exposures are first stacked along the channel axis. Then a convolution is applied followed by ReLU. We experiment with 1×1 and 3×3 kernels.

6.2.3 Late Fusion Strategy

The late fusion strategy consists in running the object detector for each of the n images independently in parallel. The final NMS stage is performed on the union set of all second stage detections. See Supplemental Material for details and comparisons with modified training losses.

6.3. Experiments

Ablations - Fusion Strategies As an ablation study to further validate the proposed fusion strategy, we conduct comparisons with the maximum pooling fusion, convolutional fusion (1×1 and 3×3) and late fusion methods described in Section 6.2. The proposed method, local cross-attention fusion, performs overall better than these four alternatives by respectively 1%, 5.7%, 6.3% and 1.2% mAP. This is confirmed by a higher AP score across most of the

considered object classes, see Table 1. Specifically, the proposed feature fusion strategy outperforms the pixel-level fusion methods in five out of the six considered object classes. These experiments confirm the effectiveness of the proposed fusion block and other fusion strategies.

Ablations - Local Cross Attention at Different Stages In addition to the default local cross-attention fusion at the end of the feature extractor, we propose fusing different exposure features at varying stages of the 28-layer ResNet variant [13]. We follow the terminology of [13], where *Conv1* (43.3 mAP) refers to the initial 7x7 convolution and *Conv2* (43.9 mAP)/*Conv3* (44.1 mAP)/*Conv4* (44.2 mAP) to the following three residual blocks of the feature extractor. Results show that performance increases when fusing at later stages, however, by trading off runtime, which we discuss in Sec. 6.4) and the Supplemental Dcoument.

Ablations - Number of Exposures Additionally we vary the number of exposures (see Table 2). We experiments using two, three and four exposures all using the proposed Local Cross Attention fusion module after *Conv4*. We chose three exposures as our default configuration (Table 1) for all other experiments as a quality/run-time trade-off.

Ablations - Local Cross Attention Fusion Module The proposed Local Cross-Attention Fusion differs in key aspects from [37]. The main design choices are experimentally validated in Table 2. a) The 2-dim SoftMax, taylored to the purpose of fusion (+1.0% compared to conventional 1D version), b) Normalized Queries (+1.9%) and c) Learning a projection matrix for the keys W^K (+0.1%). Other internal components (see Figure 4) cannot be ablated.

Quantitative and Qualitative Analysis Next, we compare the proposed method to the baseline detection methods. We report our findings in Tab. 1. These evaluations validate that the proposed neural fusion variants, which use three exposures, outperform the HDR baselines. The proposed method is overall best by 2.7% mAP and 2.3% mAP, respectively compared to *Raw HDR* and *Deep HDR*.

Fig. 5 shows qualitative results that complement the quantitative analysis. Objects in the darkest parts of the image can be missed by Raw HDR and Deep HDR methods, while the proposed Local Cross Attention Fusion method manages to detect them. We find the presence of the highest exposure capture is particularly useful in such cases. Such an instance can be seen in the first row of images in Fig. 5, where a particularly difficult to distinguish vehicle is in front of a house on the left side of the image. This vehicle is not detected by Raw HDR and Deep HDR methods, but the proposed Local Cross Attention Fusion method manages to detect it. This is also the case for a vehicle parked on the left of the image in the second row of images, whose detection escapes the Raw HDR method. Hallucinating HDR images, the Deep HDR method suffers from false negatives in this particularly dark part of the image.

Partially occluded objects are another challenge for detection in high dynamic range scenes. Occlusions can be due to the presence of other objects masking the object of interest, as is the case in the fourth row of images in Fig. 5 where a truck parked in a poorly lit area is partially occluded by a tree and a low wall. Despite this, the proposed method manages to detect this truck, whereas the Deep HDR and Raw HDR methods fail to detect it. Finally, some objects can also be occluded because they exit the camera field of view, so that only a small part of the object is visible. When combined with the fact that this small part of the object is poorly exposed because the camera must be able to properly expose a high dynamic range image, this makes the object particularly difficult to detect. This is the case, for example, with a car in the third row of Fig. 5, which disappears to the left of the image. The Deep HDR and Raw HDR methods fail here, while the Local Cross Attention Fusion method manages to detect this object which is very badly exposed and largely occluded. Finally, small objects are known to be a source of difficulty for object detectors. This difficulty is even more pronounced in a high dynamic range situation, such as the tunnel entrance visible in the last row of Fig. 5, where we can see small traffic signs at the entrance of the tunnel, which the Raw HDR and Deep HDR methods fail to detect, but which are detected correctly by the proposed Local Cross Attention Fusion method.

6.4. Runtime and Additional Model Parameters

Next, we analyze the inference runtime and the parameters that are added by the proposed feature-fusion approach.

Number of Additional Parameters For Local Cross-Attention Fusion, additional parameters are introduced by the matrices Q and W^K. This adds $d^2 + q \cdot d = 16,896$ parameters ($d = 128, q = 4$). In contrast, for Convolutional Fusion with $k \times k$ kernels. This adds $k^2 \cdot n \cdot d \cdot (d+1)$ parameters. For $k = 1$ and 3, this is respectively 49,536 and 445,824 parameters. For Maximum Pooling Fusion and Late Fusion, there are no additional parameters.

Runtime Complexity Analysis Because of its locality, our cross-attention fusion has a complexity comparable to a 1×1 convolution, and can be implemented using 2D convolutions. For instance the matrix multiplications in equations (8) and (9) on the whole image can be implemented as 1×1 convolutions. Their computational complexities are respectively $O(h \cdot w \cdot n \cdot d^2)$ and $O(h \cdot w \cdot n \cdot q \cdot d)$.

Real-Time Runtime on Jetson Orin AGX Using the lightweight ISP described in the Supplemental Material with our method, batch processing with $n = 3$ exposures runs in 17.5 ms (57 FPS) on the Nvidia Jetson AGX Orin

Figure 5. Qualitative comparison of the proposed *Local Cross-Attention Fusion* with the baseline methods *Raw HDR* and *Deep HDR* [16] on challenging scenes. Our neural fusion module recover features from separate exposure streams, where the image region is well exposed to make its decision. In contrast, the fused HDR image misses details and local contrast resulting in false negatives and false positives. Final detections in the last three columns are overlaid over all Cross-Attention exposures.

Table 3. Complexity analysis, runtime on Jetson Orin AGX (FPS) and relative detection performance compared to the proposed Local Cross Attention Fusion after *Conv4*.

Method	Point of Fusion	GFLOPS	FPS	ΔmAP
Shim et al. [36] (LDR)	N/A	44	144	-29.6
Onzon et al. [31] (LDR)	N/A	44	144	-3.1
RAW HDR	Pre-ISP	44	144	-2.7
Deep HDR [16]	Post-ISP	401	16	-2.3
Max Pooling Fusion *(ours)*	Conv4	111	57	-1.0
Local Cross Attention *(ours)*	Conv4	111	57	0.0
Conv 1 x 1 Fusion *(ours)*	Conv4	111	57	-5.7
Local Cross Attention *(ours)*	Conv3	90	70	-0.1
Conv 3 x 3 Fusion *(ours)*	Conv4	111	57	-6.3
Local Cross Attention *(ours)*	Conv2	72	88	-0.3
Late Fusion *(ours)*	2nd Stage	131	48	-1.2
Local Cross Attention *(ours)*	Conv1	51	123	-0.9

(16-bit float) using the proposed Local Cross Attention Fusion after *Conv4*. Runtimes, complexity and relative detection performance of the compared variants can be found in Table 3. Proposed variants that fuse exposure features at earlier stages achieve faster runtimes with just slight losses in downstream detection performance. Note that we run our method on a general-purpose GPU while ISPs typically run on specialized ASICs.

7. Conclusion

Outdoor scenarios are challenging for computer vision because of large dynamical ranges of luminance. Instead of a conventional HDR sensing approach with fusion in image space, we propose a neural exposure fusion that attends to the information of different LDR captures using a novel local cross attention module, allowing for fusion of the information in feature space. This feature-based fusion is embedded in an end-to-end trainable vision pipeline that jointly learns exposure control, image processing, feature extraction and detection driven by a downstream loss. Our method outperforms conventional HDR, learned, and classical auto-exposure methods on challenging automotive scenarios, when sunlight and low-lit areas are present in the same scene, where HDR fusion strategies lead to poorly exposed areas and fail to yield robust features for detection. In the future, we plan to expand our method to other vision tasks, including segmentation and optical flow, and investigate the design of the sensor itself, such as pixel layout and readout schemes — all driven by the downstream task.

Acknowledgements Work supported by AI-SEE and NRC Canada. Torc Robotics provided the data and resources.

References

[1] Yeyao Chen, Gangyi Jiang, Mei Yu, You Yang, and Yo-Sung Ho. Learning stereo high dynamic range imaging from a pair of cameras with different exposure parameters. *IEEE TCI*, 6: 1044–1058, 2020. 3

[2] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016. 3

[3] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '08*, 1997. 2, 6

[4] 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. 3

[5] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM TOG (SIGGRAPH ASIA)*, 36(6), 2017. 3

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

[7] 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

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2, 5

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

[10] Param Hanji, Fangcheng Zhong, and Rafał K Mantiuk. Noise-aware merging of high dynamic range image stacks without camera calibration. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 376–391. Springer, 2020. 6

[11] 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

[12] 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

[13] 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. 3, 4, 7

[14] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017. 3

[15] 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

[16] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36:144:1–144:12, 2017. 3, 6, 8

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

[18] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. Fhdr: Hdr image reconstruction from a single ldr image using feedback network. *arXiv preprint*, 2019. 3

[19] Jung Hee Kim, Siyeong Lee, Soyeon Jo, and Suk-Ju Kang. End-to-end differentiable learning to hdr image synthesis for multi-exposure images. *AAAI*, 2020. 3

[20] 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. 3

[21] Huei-Yung Lin and Wei-Zhe Chang. High dynamic range imaging for stereoscopic scene representation. In *ICIP*, pages 4305–4308. IEEE, 2009. 3

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

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

[26] 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. 3

[27] Julien N. P. Martel, Lorenz K. Müller, Stephen J. Carey, Piotr Dudek, and Gordon Wetzstein. Neural sensors: Learning pixel exposures for hdr imaging and video compressive sensing with programmable sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1642–1653, 2020. 3

[28] 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

[29] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020. 3

[30] Ratnajit Mukherjee, Miguel Melo, Vítor Filipe, Alan Chalmers, and Maximino Bessa. Backward compatible object detection using hdr image content. *IEEE Access*, 8: 142736–142746, 2020. 3

[31] Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2021. 3, 5, 6, 8

[32] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 3, 2017. 3

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[34] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 1, 2, 3

[35] 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. 2, 3, 4, 5

[36] 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. 6, 8

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5, 7

[38] Jian-Gang Wang and Lu-Bing Zhou. Traffic light recognition with high dynamic range imaging and deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(4): 1341–1352, 2018. 3

[39] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[40] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 3

[41] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE TIP*, 29:4308–4322, 2020. 3