Neural Exposure Fusion for High-Dynamic Range Object Detection

Emmanuel Onzon¹  Maximilian Bömer¹  Fahim Mannan¹  Felix Heide¹,²
¹Torc Robotics  ²Princeton 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.

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.

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-
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

\[
I_{HDR} = \text{ExpoFusion}(R_1, \ldots, R_n).
\]

(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 \( \phi_{scene} \). Specifically, an image \( R_j, j \in \{1, \ldots, n\} \), with exposure time \( t_j \) and gain setting \( K_j \) is

\[
R_j = \min((\phi_{scene} \cdot t_j + n_{pre}) \cdot g \cdot K_j + n_{post}, M_{white}),
\]

(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,

\[
I_{HDR} = \sum_{j=1}^{n} w_j \cdot R_j,
\]

(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

\[
(b_i, c_i, s_i)_{i \in I} = \text{ODH}(\text{ISP}_{\text{hw}}(\text{ExpoFusion}(R_1, \ldots, R_n))),
\]

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, ISP_{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

\[
(b_i, c_i, s_i)_{i \in I} = \text{ODH}\left(\text{Fusion}\left(\text{FE}(\text{ISP}(R_1)), \ldots, \text{FE}(\text{ISP}(R_n))\right)\right).
\]

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 multi-exposure HDR capture and vision pipeline where exposure control, ISP, feature extraction, 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.

![Local Cross-Attention Fusion](image)

**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 \( \alpha_{r,c} \), which is used to produce the vector at location \( (r,c) \) of the fused feature map \( f_{im} \). 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

\[
y_{j,r,c,k} = (\text{FE}(\text{ISP}(R_j)))_{r,c,k},
\]

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 \( f_{im} \) is then computed by applying a weighted sum reduction across the first dimension, that is

\[
f_{im,(r,c,k)} = \sum_{j=1}^{n} \alpha_{j,r,c} \cdot y_{j,r,c,k}.
\]

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 \( y_{r,c} \), such that for \( 1 \leq j \leq n \) and \( 1 \leq k \leq d \), \( y_{r,c,j,k} = y_{j,r,c,k} \), and apply our attention module locally to produce the \( n \)-dimensional local attention weight vector \( \alpha_{r,c} \),

\[
\alpha_{r,c} = \text{Attention}(Q, y_{r,c}),
\]
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

$$K^{(r,c)} = y_{r,c}. \bar{W}^k,$$

where $\bar{W}^k$ 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 $\alpha$ by the matrix $\tilde{\alpha}_{i,r,c}$, which is computed as

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

The matrix $\tilde{Q}$ is the normalized version of matrix $Q$, where each row of $Q$ has been divided by its $\ell^2$ 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,

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

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

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

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 $y_{r,c}$. While a projection matrix $\bar{W}^v$ 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 decomposed. 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 $I_{\text{HDR}}$ 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.
We simulate captures using 7 exposure shifts $\kappa_{\text{shift}} \in \{ -15, -10, 0, 5, 10, 15 \}$, 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 the single 20-bit raw HDR image is simulated, as explained in Section 5. This image is then used as input to the ISP.

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Point of Fusion</th>
<th>Classes</th>
<th>Bike</th>
<th>Bus &amp; Car</th>
<th>Person</th>
<th>Traffic</th>
<th>Van</th>
<th>Light</th>
<th>Sign</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shim et al. [36] (LDR)</td>
<td>N/A</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Ozuon et al. [31] (LDR)</td>
<td>N/A</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Raw HDR</td>
<td>Pre-ISP</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Debevec and Malik [3]</td>
<td>Post-ISP</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Deep HDR [16]</td>
<td>Post-ISP</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Hanji et al. (PNPE) [10]</td>
<td>Pre-ISP</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Max Pooling Fusion (ours)</td>
<td>Conv4</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Conv 1 x 1 Fusion (ours)</td>
<td>Conv4</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Conv 3 x 3 Fusion (ours)</td>
<td>Conv4</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Late Fusion (ours)</td>
<td>2nd Stage</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
<tr>
<td>Local Cross Attention (ours)</td>
<td>Conv4</td>
<td>Bike, Bus &amp; Car, Person, Traffic, Van</td>
<td>25.7</td>
<td>15.8</td>
<td>41.8</td>
<td>41.8</td>
<td>53.8</td>
<td>42.3</td>
<td>1.2%</td>
<td>43.2</td>
</tr>
</tbody>
</table>

#### 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,

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

#### 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 $1 \times 1$ and $3 \times 3$ 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 $(1 \times 1 \times 3 \times 3)$ 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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Point of Fusion</th>
<th>Bike</th>
<th>Bus &amp; Car</th>
<th>Person</th>
<th>Traffic</th>
<th>Van</th>
<th>Light</th>
<th>Sign</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoftMax 1D</td>
<td>Conv4</td>
<td>26.7</td>
<td>16.7</td>
<td>73.6</td>
<td>45.9</td>
<td>42.5</td>
<td>53.6</td>
<td>44.1</td>
<td>43.2</td>
</tr>
<tr>
<td>Skip Normalize Q</td>
<td>Conv4</td>
<td>25.7</td>
<td>15.8</td>
<td>71.4</td>
<td>45.3</td>
<td>41.8</td>
<td>53.8</td>
<td>44.1</td>
<td>43.2</td>
</tr>
<tr>
<td>Skip Multiply W(K)</td>
<td>Conv4</td>
<td>26.5</td>
<td>17.2</td>
<td>74.0</td>
<td>46.9</td>
<td>44.2</td>
<td>55.9</td>
<td>44.1</td>
<td>43.2</td>
</tr>
<tr>
<td>LCA 2 Exposures</td>
<td>Conv4</td>
<td>27.1</td>
<td>16.5</td>
<td>74.1</td>
<td>46.7</td>
<td>42.3</td>
<td>55.7</td>
<td>43.7</td>
<td>43.2</td>
</tr>
<tr>
<td>LCA 4 Exposures</td>
<td>Conv4</td>
<td>26.7</td>
<td>16.7</td>
<td>74.4</td>
<td>47.9</td>
<td>44.3</td>
<td>56.2</td>
<td>44.4</td>
<td>43.2</td>
</tr>
</tbody>
</table>
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 Document.

**Ablations - Number of Exposures**

Additionally we vary the number of exposures (see Table 2). We experiment 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, tailored 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 the proposed method is overall best by 2.7% mAP and 2.3% mAP, respectively compared 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 the baseline detection methods.

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.

**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.
Deep HDR Raw HDR Cross-Attention Exposure 1 Cross-Attention Exposure 2 Cross-Attention Exposure 3

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Point of Fusion</th>
<th>GFLOPS</th>
<th>FPS</th>
<th>ΔmAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shim et al. [36] (LDR)</td>
<td>N/A</td>
<td>44</td>
<td>144</td>
<td>-29.6</td>
</tr>
<tr>
<td>Onzon et al. [31] (LDR)</td>
<td>N/A</td>
<td>44</td>
<td>144</td>
<td>-3.1</td>
</tr>
<tr>
<td>RAW HDR</td>
<td>Pre-ISP</td>
<td>41</td>
<td>144</td>
<td>-2.7</td>
</tr>
<tr>
<td>Deep HDR [16]</td>
<td>Post-ISP</td>
<td>401</td>
<td>16</td>
<td>-2.3</td>
</tr>
<tr>
<td>Max Pooling Fusion (ours)</td>
<td>Conv4</td>
<td>111</td>
<td>57</td>
<td>-1.0</td>
</tr>
<tr>
<td>Local Cross Attention (ours)</td>
<td>Conv4</td>
<td>111</td>
<td>57</td>
<td>-5.7</td>
</tr>
<tr>
<td>Conv 1 x 1 Fusion (ours)</td>
<td>Conv4</td>
<td>111</td>
<td>57</td>
<td>-6.3</td>
</tr>
<tr>
<td>Local Cross Attention (ours)</td>
<td>Conv3</td>
<td>90</td>
<td>70</td>
<td>-0.1</td>
</tr>
<tr>
<td>Conv 3 x 3 Fusion (ours)</td>
<td>Conv4</td>
<td>111</td>
<td>57</td>
<td>-6.3</td>
</tr>
<tr>
<td>Local Cross Attention (ours)</td>
<td>Conv2</td>
<td>72</td>
<td>88</td>
<td>-0.3</td>
</tr>
<tr>
<td>Late Fusion (ours)</td>
<td>2nd Stage</td>
<td>131</td>
<td>48</td>
<td>-1.2</td>
</tr>
<tr>
<td>Local Cross Attention (ours)</td>
<td>Conv1</td>
<td>51</td>
<td>123</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

(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


[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


