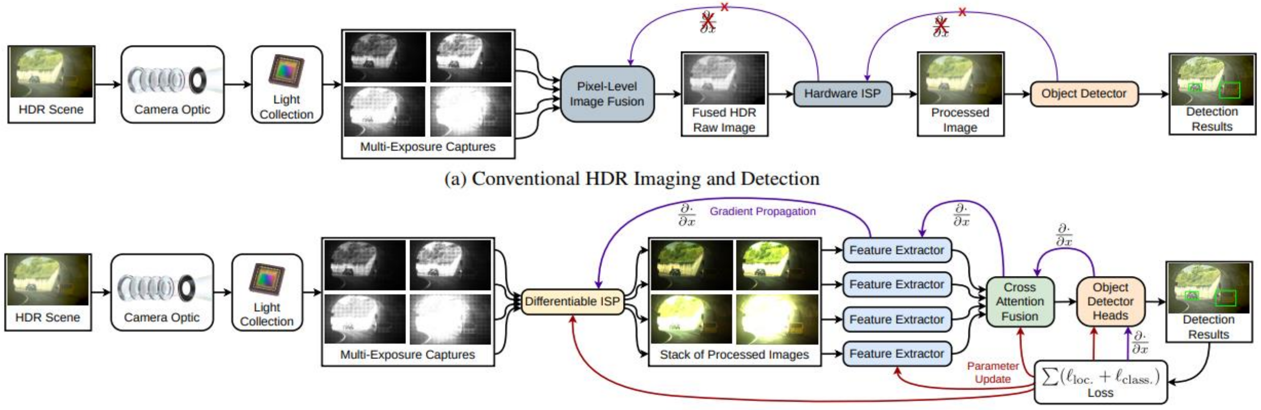
Neural Exposure Fusion for High-Dynamic Range Object Detection

What is the problem?

The challenge is of object detection in high dynamic range (HDR) outdoor scenes, especially under rapidly changing illumination conditions, such as driving scenarios with varying lighting. The computer vision system must also be able to adapt, for example, when the vision system, or when large objects in the environment move quickly.

What has been done earlier?

Traditional methods rely on HDR sensors that capture multiple low dynamic range (LDR) exposures, which are fused into a single HDR image using a hardware image signal processor (ISP). However, this fusion is designed primarily for human perception and does not retain all the necessary details for downstream computer vision tasks like object detection. As a result, important information is often lost during the fusion process, leading to suboptimal performance in detection tasks.



(a) Conventional HDR Imaging and Detection

(b) Proposed Neural Exposure Fusion for Detection

Harshit Goel, B421023

What are the remaining challenges?

Complexity and Efficiency: While the proposed method offers improved accuracy, it also introduces additional computational complexity due to the fusion of features across multiple exposures. Real-time performance, especially in resource-constrained environments like autonomous vehicles, remains a challenge. Achieving a balance between model complexity and runtime performance is crucial, particularly when deploying on specialized hardware.

Dynamic Lighting Conditions: Rapidly changing illumination conditions, such as moving from indoor to outdoor environments or facing strong backlights, still pose difficulties for even advanced HDR systems. While the proposed feature fusion method improves performance, handling extreme lighting variations with higher efficiency is an ongoing challenge.

Occlusion and Small Object Detection: The method struggles with partially occluded objects and small objects in HDR scenes. These cases are difficult to detect when they are both poorly exposed and occluded by other objects. Although the attention mechanism improves detection, further enhancement is needed to fully address these issues.

What novel solution proposed by the authors to solve the problem?

Task-Driven Feature Fusion:

- Instead of fusing multiple low dynamic range (LDR) exposures into a single HDR image, the authors propose to fuse features from different exposures directly in the feature space. This approach allows the fusion process to be optimized for the specific task of object detection, retaining more task-relevant information that is often lost in image-space fusion.
- The method is trained end-to-end, with supervision coming from the object detection loss function, ensuring that the fused features are directly optimized for the detection task.

Local Cross-Attention Mechanism:

- The authors introduce a **local cross-attention fusion module** that allows the network to weigh and combine information from different exposures based on the relevance of the content for object detection.
- This attention mechanism operates on features at different spatial locations, dynamically determining which exposure contains the most useful information for the detection task. For example, it may prioritize lower exposure for bright areas and higher exposure for darker areas.
- The attention maps generated by this module provide a fine-grained, localized weighting of the exposures, ensuring that the most semantically relevant and well-exposed features are used in each part of the image.

End-to-End Differentiable Pipeline:

- The authors integrate the feature fusion into a fully differentiable vision pipeline, which includes exposure control, image signal processing (ISP), feature extraction, and object detection. All components of this pipeline are trained together in a task-specific manner, which contrasts with traditional pipelines where each component (sensor, ISP, and vision model) is trained or designed independently.
- This joint optimization leads to improved performance because the entire system is trained to maximize object detection accuracy, rather than focusing on producing visually appealing images.