6-2024

# Criticality aware canvas-based visual perception at the edge

Ila GOKARN
*Singapore Management University*, ingokarn.2019@phdcs.smu.edu.sg

# Criticality Aware Canvas-based Visual Perception at the Edge

Ila Gokarn
Singapore Management University
Singapore
ingokarn.2019@phdcs.smu.edu.sg

## ABSTRACT

Efficient and effective machine perception remains a formidable challenge in sustaining high fidelity and high throughput of perception tasks on affordable edge devices. This is especially due to the continuing increase in resolution of sensor streams (e.g., video input streams generated by 4K/8K cameras and neuromorphic event cameras that produce ≥ 10 MEvents/second) and computational complexity of Deep Neural Network (DNN) models, which overwhelms edge platforms, adversely impacting machine perception efficiency. Given the insufficiency of the available computation resources, a question then arises on whether selected regions/components of the perception task can be prioritized (and executed preferentially) to achieve highest task fidelity while adhering to the resource budget. This extended abstract explores the paradigm of *Canvas-based Processing* and criticality-awareness in the context of multi-sensor machine perception pipelines on resource-constrained platforms, in guiding perception pipelines and systems on "what" to pay attention to in the sensing field and "when", to maximize overall perception fidelity under computational constraints and moderate the processing throughput-vs-accuracy trade-off.

## CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**.

## KEYWORDS

Edge AI, Machine Perception, Canvas-based Processing

**ACM Reference Format:**
Ila Gokarn. 2024. Criticality Aware Canvas-based Visual Perception at the Edge. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3643832.3661386

## 1 INTRODUCTION

Visual machine perception is a fundamental tenet of cyber-physical systems that enables applications such as augmented reality (AR/VR), autonomous cars or drones, and assistive robots. Real-time sensing and accurate sense-making is crucial for these applications to execute sub-tasks such as object detection, localization, and mapping, yielding the ability to perceive, navigate, and interact with their physical environment. State of the art visual perception pipelines rely on the processing of video streams from commodity cameras, and often leverage multiple cameras concurrently to achieve fine-grained, multi-perspective, or wide-range visual perception [2]. In parallel, advances in edge computing have enabled deep learning on resource constrained pervasive edge devices to consume such camera streams and distill the perception of physical phenomena into actionable knowledge [2]. However, simultaneous advancements in (i) newer and more complex visual Deep Learning Networks (DNNs) that impose higher memory-latency complexity, and (ii) higher resolution cameras (i.e. 4k/8k cameras and 1MEvent/second neuromorphic event cameras) that generate data at high velocities and resolutions, make it challenging to guarantee real-time, efficient, and accurate visual perception over multiple such camera streams using a *single* resource-constrained edge device.

To address this challenge, we introduce the paradigm of *Criticality Aware Canvas-based Processing* and explore multi-sensor visual perception pipelines on resource constrained embedded platforms, with a focus on improving end-to-end system efficiency. Inspired by the concept of *attention* from human psychology, we consider criticality-awareness as the selective concentration of limited computation resources on a smaller subset of stimuli among multiple perceivable stimuli [7]. While the concept of *attention* has been adapted by the deep learning community to make DNNs dynamically focus on relevant parts of the input data, we make a key distinction between attention mechanisms and criticality-awareness in that we focus on fine-tuning the *entire* edge system and application across *multiple inputs*, not the structure of the DNN itself. Canvas-based Processing involves the extraction of critical stimuli or Regions of Interest (RoI) from multiple concurrent camera streams, to create a spatial and temporal degree of freedom for each stimuli from its original video source, thus allowing the system to spatiotemporally multiplex limited computation resources to selected stimuli for DNN inference. Fine-tuning of such a system system can be borne out of (i) *what* subset of stimuli the system selects for fine-grained perception of downstream DNN task execution [3] (ii) *when* the system re-selects already perceived objects to monitor changes in relative criticality to the perception task [1] and (iii) *how* the resource-constrained edge GPU schedules critical stimuli for downstream DNN inference [4–6]. We hypothesize that, in general, continuous recognition is not strictly necessary for machine perception given as how human perception follows similar principles with reasonable efficiency. We believe that by fine-tuning the system's understanding of the different stimuli across multiple sensors, visual perception tasks operating on a single edge device may be afforded with higher system throughput while maintaining DNN task accuracy. In Section 2, we describe our approaches to address these goals, highlighting our main findings and scope of future work.


MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06.
https://doi.org/10.1145/3643832.3661386

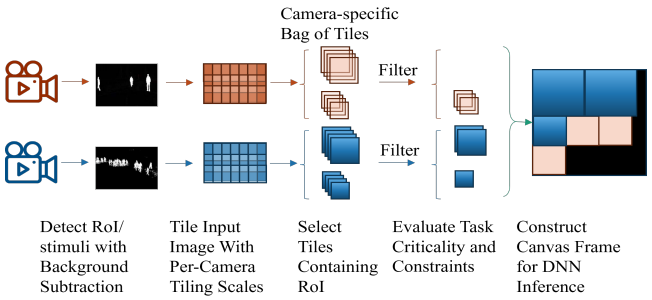


**Figure 1: Overview of Criticality Aware Canvas-based Processing over multiple input streams at a single edge GPU**

## 2 THE CANVAS-BASED PROCESSING PARADIGM

In general, we define a *Canvas* as the maximum size of an input frame that a DNN operating on an edge GPU device can consume to yield an inference throughput above the desired real-time processing threshold [3]. This blank canvas frame essentially acts as a medium, populating a composite image constructed of selected stimuli for downstream DNN inference such as object detection, illustrated in Figure 1.

**Introducing a Spatial Degree of Freedom** To address the first notion of *what* subset of stimuli the system must select for fine-grained perception, we first introduce a spatial degree of freedom to the stimuli (hereafter referred to as Regions of Interest (RoI)) [3]. We identify RoI in the sensing field using motion-based background subtraction and decompose each input frame from $M$ concurrent camera streams into a "a bag of tiles" or regions of the frame, capturing the detected RoI at different camera-specific scales or "levels of zoom". As multiple tiles at different scales may capture the same object, we filter and select only those tiles that capture all the RoI at their appropriate scale, ultimately spatially multiplexing these filtered tiles via Inverse 2D Bin Packing onto the limited volume of the canvas frame for DNN inference. We observe that such spatial multiplexing yields substantial gains in the throughput-accuracy trade-off where a single Jetson TX2 device executing an edge-scale DNN model is capable of packing RoI from $M = 6$ concurrent camera streams to achieve 4.75× (475%) higher throughput (23 FPS *per camera*, cumulatively 138FPS) with $\leq 1\%$ accuracy loss, compared to a First Come First Serve (FCFS) processing paradigm.

**Exploring a Temporal Degree of Freedom** To address the second notion of *when* the system re-selects already perceived objects to monitor changes in its physical environment, we introduce a temporal degree of freedom using a novel age utility-based weighted scheduler to preferentially deprioritize recently seen unique objects for inclusion on a canvas frame constructed over multiple spatially-overlapped cameras [1]. To balance between accurate perception of the state of the physical world and delay induced by the age metric in the processing pipeline, we selectively discard multiple instances of RoI over time to optimize the novel streaming accuracy metric [8] which accounts for localisation latency in addition to object detection accuracy. For an exemplar traffic surveillance application, such spatiotemporal multiplexing of selected RoI on a canvas frame enables the concurrent processing of up to $M = 25$ cameras on a single Jetson TX2 device with a 66.6% increase in streaming accuracy and a simultaneous 18× (1800%) gain in cumulative processing throughput, compared to competitive baselines.

**Algorithms for Scheduling Canvas Construction** In exploring *how* the resource-constrained edge GPU must schedule critical stimuli for downstream DNN inference, we derive spatiotemporal schedulability bounds and Earliest Deadline First based bin-packing algorithms for (i) RoI that are quantized to fixed/known dimensions prior to inclusion on the canvas [4, 5] and (ii) unquantized dynamically resized RoI [6]. We show that while the best order in which RoI must be considered for inclusion on the canvas frame can be dramatically different from the best order in which the RoI must be considered to meet schedulability deadlines, efficient packing yields effective available capacity on the canvas frame, and ultimately the ability to better meet schedulability deadlines.

**Canvas Construction Methods and Evaluation** We explore a number of canvas construction methods (i) Binary Quadtree based Inverse 2D bin packing [3, 6], (ii) Genetic Algorithm based differential evolution to squeeze-pack dynamically sized RoI onto a canvas frame [1], (iii) Quantized Binary Quadtree based Inverse 2D Bin-packing [4, 5] to adapt to different {spatial, temporal, spatiotemporal} packing objectives of each processing pipeline. We also evaluate these mechanisms in real-world wireless camera deployments to characterize the gains in throughput-vs-accuracy under varying bandwidth-constrained environmental conditions [1].

## 3 FUTURE WORK AND OUTLOOK

We show that the criticality-aware canvas-based processing paradigm is a powerful system optimization technique that pushes the envelope on achievable throughput-vs-accuracy trade-offs on resource constrained edge devices, enabling the processing of a greater number of cameras/sensors on a single edge GPU device and drastically reducing infrastructure costs with little to no loss in perceptual fidelity. We intend to extend this body of work to address (i) object arrival rates which can pre-determine canvas utilisation bounds to admit a greater number of cameras in the "wait times" (ii) efficient frame transmission mechanisms that complement the extraction of stimuli/RoI from the input frame and are robust to varying wireless latency and bandwidth characteristics, (iii) workload adaptive canvas construction mechanisms that opportunistically yield smaller volumes of canvas frames to gain average processing throughput, and (iv) how stimuli from multiple other sensors such as novel neuromorphic event-based cameras might impact relative criticality of different objects/events to the perception task.

## ACKNOWLEDGMENTS

This work was supported by National Research Foundation, Singapore under its NRF Investigatorship grant (NRF-NRFI05-2019-0007), and in part by The Boeing Company, IBM (IIDAI), ARL W911NF-17-2-0196, and NSF CNS 20-38817. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

## REFERENCES


[1] Ila Gokarn, Yigong Hu, Tarek Abdelzaher, and Archan Misra. 2024. JIGSAW: Edge-based Streaming Perception over Spatially Overlapped Multi-Camera Deployments. In *Proceedings of the IEEE International Conference on Multimedia Expo (ICME) (to appear)*.

[2] Ila Gokarn, Kasthuri Jayarajah, and Archan Misra. 2023. Lightweight Collaborative Perception at the Edge. In *Artificial Intelligence for Edge Computing*. Springer, 265–296.

[3] Ila Gokarn, Hemanth Sabbella, Yigong Hu, Tarek Abdelzaher, and Archan Misra. 2023. MOSAIC: Spatially-multiplexed edge AI optimization over multiple concurrent video sensing streams. In *Proceedings of the 14th Conference on ACM Multimedia Systems*. 278–288.

[4] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, and Tarek Abdelzaher. 2023. Underprovisioned gpus: On sufficient capacity for real-time mission-critical perception. In *2023 32nd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–10.

[5] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, and Tarek Abdelzaher. 2023. Work-in-Progress: Algorithms for Canvas-Based Attention Scheduling with Resizing. In *2023 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 435–438.

[6] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, and Tarek Abdelzaher. 2024. Algorithms for Canvas-Based Attention Scheduling with Resizing. In *Proceedings of 30th IEEE Real-Time and Embedded Technology and Applications Symposium (to appear)*. IEEE.

[7] Christof Koch and Shimon Ullman. 1987. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*. Springer, 115–141.

[8] Mengtian Li, Yuxiong Wang, and Deva Ramanan. 2020. Towards Streaming Perception. *ECCV* (2020).


## ABOUT THE AUTHOR

**Ila Gokarn** is currently pursuing her PhD in Computer Science at Singapore Management University having obtained her BSc in Information Systems from the same university in 2015. Her research work lies in pervasive sensing and edge computing paradigms, published in research papers and a book chapter.