

R-TOD: Real-Time Object Detector with Minimized End-to-End Delay for Autonomous Driving

Wonseok Jang¹, Hansaem Jeong¹, Kyungtae Kang², Nikil Dutt³, and Jong-Chan Kim^{1,4}

¹Graduate School of Automotive Engineering, Kookmin University, Korea

²Department of Computer Science and Engineering, Hanyang University, Korea

³Department of Computer Science, University of California, Irvine, USA

⁴Department of Automobile and IT Convergence, Kookmin University, Korea

xkqhdl@kookmin.ac.kr, saem28@kookmin.ac.kr, ktkang@hanyang.ac.kr, dutt@uci.edu, jongchank@kookmin.ac.kr

Abstract—For realizing safe autonomous driving, the end-to-end delays of real-time object detection systems should be thoroughly analyzed and minimized. However, despite recent development of neural networks with minimized inference delays, surprisingly little attention has been paid to their end-to-end delays from an object's appearance until its detection is reported. With this motivation, this paper aims to provide more comprehensive understanding of the end-to-end delay, through which precise best- and worst-case delay predictions are formulated, and three optimization methods are implemented: (i) on-demand capture, (ii) zero-slack pipeline, and (iii) contention-free pipeline. Our experimental results show a 76% reduction in the end-to-end delay of Darknet YOLO (You Only Look Once) v3 (from 1070 ms to 261 ms), thereby demonstrating the great potential of exploiting the end-to-end delay analysis for autonomous driving. Furthermore, as we only modify the system architecture and do not change the neural network architecture itself, our approach incurs no penalty on the detection accuracy.

I. INTRODUCTION

For the realization of safe autonomous driving, a vehicle's camera-based object detection system should detect hazardous on-road obstacles as quickly as possible to reduce the risk of collision. Thus, delays between the appearance of an object and its detection should be thoroughly analyzed and optimized. Since most recent object detectors are based on *deep neural networks* (DNNs), lots of studies have attempted to reduce DNN *inference delays* by developing lightweight neural networks [1], [2] or hardware acceleration techniques [3], [4]. However, surprisingly little attention has been paid to the analysis and optimization of the *end-to-end delays* of real-time object detection systems, which include not only inference delays but also other pre- and post-processing delays such as resizing images and displaying results. In this paper, the end-to-end delay is defined by from the physical appearance of an object (e.g., change of a traffic signal or sudden pedestrian crossing) until its first detection is reported.

In autonomous driving systems, the frame rates, as well as the end-to-end delays, are commonly considered as the most important performance metrics [5]. In this regard, Fig. 1 presents our preliminary measurement results of these metrics using three different versions of Darknet YOLO (You Only Look Once) object detectors [6]–[9] on three Nvidia hardware platforms. For ease of comparison, cycle times (i.e., inverse of frame rates) are presented instead. The figure indicates

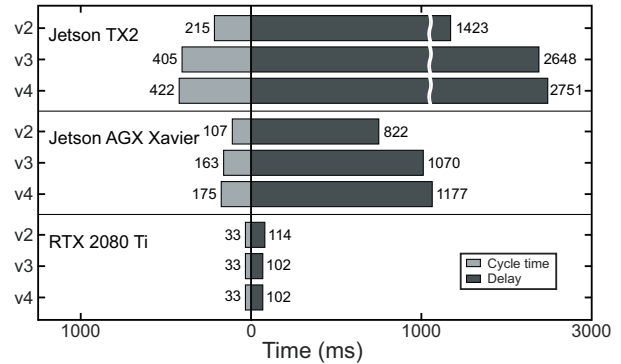


Fig. 1: Average end-to-end delays and cycle times (i.e., inverse of frame rates) of Darknet YOLO (v2, v3, and v4) object detectors on three different hardware platforms.

that different combinations of DNNs and hardware platforms exhibit significantly varying performance. Another important observation highlighted here is the *imbalance between delays and cycle times*. For example, consider the case of Jetson AGX Xavier and YOLO v3: because its cycle time is 163 ms, we naively expect the delay to have a similar value; unexpectedly, the actual measured value is 1070 ms that is around six times the cycle time. It is apparent that the end-to-end delays are excessive in relation to the intrinsic cycle times, meaning that even if we are able to obtain object detection results with sufficient frequency, they still may suffer from significant time lags to the real world. Moreover, we found similar trends in other state-of-the-art object detection systems including Detectron2 [10] and Darkflow [11]. With this observation, we pose the following research questions:

- What is the reason for the serious time lags compared with the cycle times of object detection systems?
- How can we minimize the end-to-end delays of object detection systems, and by how much?

To answer the first question, we thoroughly analyzed the internal architecture of the Darknet YOLO object detection system and measured every component that contributed to the end-to-end delay on our experimental platforms. From this investigation, we found that (i) Darknet YOLO relies on the OpenCV library for capturing real-time images from cameras,

which manages a queue for buffering image frames; and (ii) it employs a three-stage multithreaded pipeline architecture for exploiting its *pipeline parallelism* [12], [13]. After carefully analyzing the internal architecture of Darknet YOLO, we have concluded that the reason behind the serious time lags is threefold: (i) there is an unnecessary queuing delay caused by the difference between the frame rate of the camera and that of the object detector; (ii) the lengths of the three pipeline stages are much too different causing significant idle times for shorter ones; and (iii) concurrently running GPU (Graphics Processing Unit) kernels and CPU (Central Processing Unit) threads incurs significant contention for shared memory bandwidth, leading to increased execution times, especially when an integrated GPU is used.

To answer the second question, while minimizing the end-to-end delays of the Darknet YOLO object detection system, we developed the following three techniques and evaluated them on an Nvidia Jetson AGX Xavier. Note that the numbers below are for YOLO v3 in particular:

(i) On-demand capture. We removed unnecessary queuing delays by applying an *on-demand capture* method that does not use a queue between cameras and object detectors. Instead, an image frame is captured only when it is needed, thereby reducing the average delay from 1070 ms to 383 ms (i.e., a 64% reduction).

(ii) Zero-slack pipeline. We minimized pipeline idle times by postponing the beginning of the first pipeline stage as much as possible within its pipeline cycle such that the next cycle begins right after the end of the first stage. On average, this *zero-slack pipeline* saved an extra 73 ms, and together with on-demand capture, it reduced the overall delay to 310 ms.

(iii) Contention-free pipeline. To minimize the memory bandwidth contention between concurrently running GPU kernels and CPU threads, we applied synchronized executions between them, thereby achieving temporal isolation of GPU kernel executions from CPU workloads. By applying this *contention-free pipeline*, the delay was even further reduced to 261 ms, which is about a 76% reduction compared with the original one. Furthermore, to evaluate the tail behavior of the optimized system, the 99th percentile delays were compared, which showed an end-to-end delay reduction of about 67%.

The contributions of this study can be summarized as:

- We present an in-depth analysis of the Darknet YOLO object detection system and our end-to-end delay analysis framework, and identify three problems with significant adverse impacts on its end-to-end delays.
- We propose a modified OpenCV library that supports *on-demand capture* that significantly reduces the end-to-end delays of various state-of-the-art object detectors without any executable file modification.
- We present *zero-slack* and *contention-free* pipeline architectures that alleviate the delays caused by imbalanced pipeline stages and memory bandwidth contention, respectively.
- The above techniques are implemented and evaluated on an Nvidia Jetson AGX Xavier CPU-GPU heterogeneous

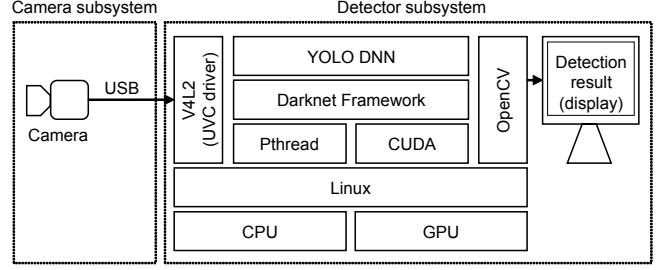


Fig. 2: Overall hardware-software architecture.

computing platform that shows 76% average and 67% 99th percentile delay reductions for YOLO v3, with no loss in detection quality.

II. BACKGROUND AND PROBLEM DESCRIPTION

A. Overall Hardware-Software Architecture

Fig. 2 depicts our reference hardware-software architecture considered for object detection systems, which consists of (i) a camera subsystem and (ii) a detector subsystem. In the camera subsystem, we consider a single camera that is directly attached to the detector subsystem using a USB (Universal Serial Bus) connector. For the detector subsystem, a CPU-GPU heterogeneous computing platform is used. The Linux operating system is used to run the Darknet neural network framework based on the pthread library and the CUDA framework for managing CPU threads and GPU kernels, respectively. In the Darknet framework, we use pre-trained YOLO v2, v3, and v4 DNNs. As an interface between the two subsystems, the Video For Linux Two (V4L2)-based USB camera driver is used. To visualize object detection results, the OpenCV library is used to show bounding boxes around the detected objects. The remainder of this section briefly summarizes details regarding Darknet, YOLO, V4L2, and OpenCV.

Darknet [6] is an open source neural network framework developed as the baseline software architecture for the YOLO object detector. Darknet can be used for training a custom neural network as well as for an inference engine with pre-trained neural networks. Darknet can utilize GPUs as well as CPUs when running neural networks. Because it is written purely in C and CUDA with minimal dependencies, it is easy to port from a high-end workstation to a lightweight embedded system. Darknet supports the most commonly used neural network operators, such as convolution, pooling, and various activation functions. Thus, a neural network configuration file with layer-by-layer descriptions and a weights file are sufficient when implementing a neural network inference system.

YOLO is one of the most famous single-stage object detectors based on the *convolutional neural network* (CNN) architecture. Because of its high speed and accuracy, it is widely used for implementing real-time object detection systems for autonomous driving [5], [14], [15]. YOLO has evolved from its original version [16] to more recent versions (v2, v3, and v4) [7]–[9] in which its object detection accuracy has been continually increased. Pre-trained neural network configuration files and weights files are available for each

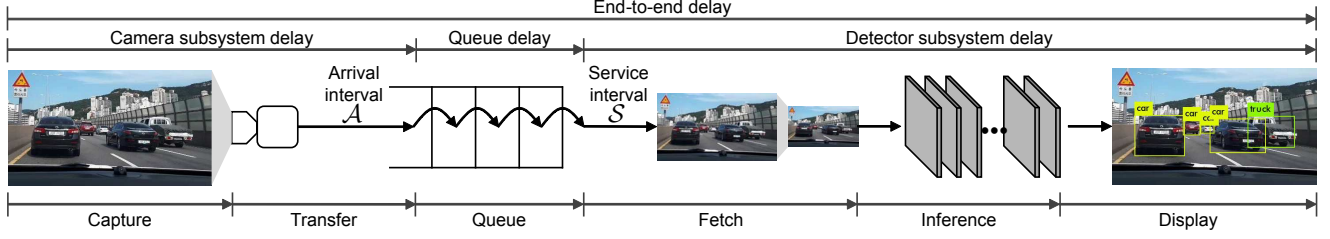


Fig. 3: Overall flow and associated delays in an object detection system.

version; these are based on open datasets such as PASCAL VOC [17] and MS COCO [18].

V4L2 [19] is a device driver framework used to process real-time video images in Linux. For example, most USB cameras are supported by the UVC (USB Video Class) driver that relies on the V4L2 framework, especially when using its userspace application programming interface (API). The V4L2 API represents a camera as a character device file such that userspace applications can open it to fetch a sequence of images through the `ioctl` system call. Furthermore, users using the V4L2 API can also control camera parameters including the frame rate, pixel format, and exposure.

OpenCV [20] is an open source computer vision library that provides various off-the-shelf computer vision algorithms. Recently, OpenCV has been used as the common infrastructure for various computer vision applications in the industry. In object detection systems, it is typically used for visualizing object detection results by drawing bounding boxes around each detected object and displaying the resulting images. Further, most object detectors, including Darknet, use the OpenCV C++ wrapper library, known as *VideoCapture*, instead of directly using the V4L2 API for capturing camera images.

B. Problem Description

Based on the hardware-software architecture described above, this study has two objectives: (i) to derive a timing model of object detection systems by analyzing the internal architecture of Darknet YOLO (Section III) and (ii) to present an optimal application architecture for object detection systems in terms of minimizing end-to-end delays (Section IV) and evaluate it with the Darknet YOLO object detection system (Section VI).

III. END-TO-END DELAY ANALYSIS

Note: Our notational conventions are as follows: (i) Roman capitals for constants (e.g., F), (ii) calligraphic capitals for probability distributions (e.g., \mathcal{A}), (iii) Roman lower-case letters for random variables (e.g., d_{capt}), and (iv) Roman lower-case letters with the superscripts min or max for the minimums or maximums of random variables (e.g., d_{capt}^{max}).

A. Holistic Approach for the End-to-End Delay Analysis

Fig. 3 shows the end-to-end flow of data in an object detection system, along with the associated delays, that comprises a camera subsystem, a detector subsystem, and a V4L2 camera driver queue (henceforth simply referred to as “queue”) for buffering images. The queue is located inside the detector

subsystem; however, for ease of explanation, we will regard it as being between the two subsystems. This entire system can be modeled as a single-server queue of size Q . Throughout the analysis, we use the distribution \mathcal{A} of image arrival intervals at the queue and the distribution \mathcal{S} of object detection service intervals, where the service interval is defined as the time between two consecutive requests for retrieving images from the queue (i.e., beginning of object detection cycles) by the detector subsystem. These distributions are discussed later in this section.

Our analysis of the end-to-end delay can now be summarized as follows:

- **Camera subsystem.** It incurs a delay defined by the time from the first appearance of an object to the queuing of its first image. We examine the components of this delay to formulate \mathcal{A} .
- **Detector subsystem.** It incurs a delay defined by the time from the retrieval of an image from the queue to the appearance of the detection result on the display. We examine the components of this delay to formulate \mathcal{S} .
- **Queue.** It incurs a delay that can be determined from \mathcal{A} and \mathcal{S} for a given queue size Q .

In Fig. 3, six different delay components are present: (i) capture, (ii) transfer, (iii) queue, (iv) fetch, (v) inference, and (vi) display; these account for the end-to-end delay. Note that the inference delay is only one among many delays, thus necessitating a holistic approach for analyzing and minimizing end-to-end delays. Moreover, although these delay components appear to be isolated from each other, some are interrelated, further indicating the need for a holistic approach.

B. Camera Subsystem

Images are captured by the camera (i.e., capture instant) at a frame rate of F , so the camera’s *cycle time* C between captures is $1/F$. At first glance, a newly appeared object seems to have to wait at most C for the next capture instant. However, the object may experience a much longer delay before it is captured, that is, the *capture delay* denoted by d_{capt} . Because it is intertwined with the detector subsystem’s service intervals, an exact formulation will be presented later in this section.

The amount of data in an image frame is determined by its resolution ($X \times Y$) and pixel format. Non-compressing pixel formats such as YUYV and RGBA have a fixed bits-per-pixel (bpp) value P . In contrast, compressing pixel formats such as H264 and MJPEG may vary the frame size depending on the image content. We assume non-compressing pixel formats

TABLE I: USB bandwidth reservations.

Resolution (X × Y)	F (fps)	B (bytes)	M (μframes)	Reserved (Required) bandwidth (MB/s)
320 × 240	20	512	32	4 (3)
	30	800	32	6 (4)
640 × 480	20	1984	32	15 (12)
	30	2688	32	21 (18)

because compressing pixel formats are non-deterministic and lossy, and they require an additional decryption process. Thus, assuming a fixed bpp, an image frame's size, denoted by I , can be calculated as $I = (X \cdot Y \cdot P)/8$ bytes.

Each captured image is transferred to the detector subsystem through a USB connection in the isochronous transfer mode, which provides a known latency by bandwidth reservation [21]. The camera driver interacts with the camera to determine the available bandwidth, which is expressed as the number of bytes, denoted by B , available during each USB microframe¹ of length $U=125 \mu s$. A USB driver uses a buffer to group a *USB request block* (URB) of M microframes (i.e., handling granularity). Table I lists the example bandwidth reservations measured in our experimental platform using the YUYV pixel format with $P = 16$.

Next, consider the *transfer delay*, denoted by d_{tran} , which is the time between the capture of an image and its addition to the queue. This can be estimated by counting the number of required microframes from given I , B , and U , as follows:

$$d_{tran} = \left(\left\lceil \frac{I}{B} \right\rceil + 2 \right) U + j, \quad (1)$$

where j is the jitter caused by (i) waiting for the start of the next microframe after a capture (U in the worst case), and (ii) waiting for the current URB to be grouped ($(M-1) \cdot U$ in the worst case). We denote the former as *microframe jitter* and the latter as *URB buffering jitter*, respectively. By summing the two, j can have values in the range $0 \leq j < M \cdot U$. Note that the constant 2 represents extra microframes containing protocol overheads.

Fig. 4 illustrates the transfer delay while taking this jitter into account. It indicates that the transfer delay is determined as the multiples of a URB buffering period of M microframes, because the camera driver only checks image arrivals at that interval. Note that $M \cdot U = 32 \cdot 125 \mu s = 4 \text{ ms}$ is true for every row of Table I, although it can vary for different cameras.

An important consequence of this approach based on URB buffering is that the image arrival intervals are adjusted to the multiples of the *URB buffering period* ($M \cdot U$) (note that their expectation approximates to the constant camera cycle time C). Thus, \mathcal{A} becomes a discrete probability distribution with (usually) two values a^{min} and a^{max} , and the probability that the image arrival interval a equals one of these values is, respectively, defined as follows:

$$p_1 = \Pr \{a = a^{min}\} \text{ and } p_2 = \Pr \{a = a^{max}\}, \quad (2)$$

¹USB manages time in units called “frames” of 1 ms. At a high speed, each 1 ms frame is divided into eight “microframes” of 125 μs .

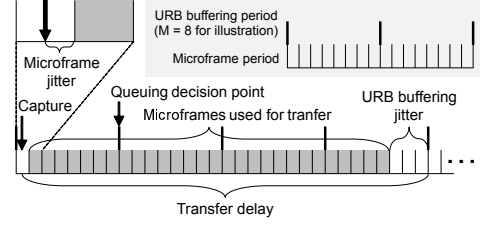


Fig. 4: Transfer delay.

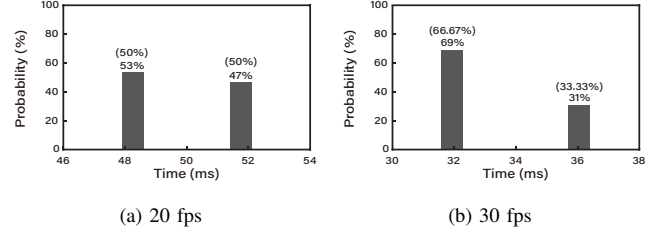


Fig. 5: Image arrival interval distribution (\mathcal{A}) with varying camera frame rates. Probabilities in parentheses are estimates; measured probabilities are written below the estimates.

where

$$a^{min} = \left\lfloor \frac{C}{M \cdot U} \right\rfloor (M \cdot U) \text{ and } a^{max} = \left\lceil \frac{C}{M \cdot U} \right\rceil (M \cdot U),$$

which represent the closest multiple of $M \cdot U$ to the floor and ceiling of C , respectively. The expected value of a ($E(a)$) can then be approximated to C , i.e., $a^{min}p_1 + a^{max}p_2 \approx C$. Since $p_1 + p_2 = 1$, the probability distribution can be derived from given C , M , and U . Fig. 5 shows the measured arrival interval distributions at two different camera frame rates: 20 fps and 30 fps. Note that the results closely match our estimates, with minor discrepancies caused by limitations in the camera hardware with respect to the exact frame rate.

When an image frame arrives at the queue, it is either accepted by the queue or dropped, where some droppings are inevitable in case the image arrivals are faster than the retrievals. In reality, the queuing decision is made even before it fully arrives. From our protocol analysis, we found that the decision is made when the first URB buffering period containing the image frame ends, and we call the instant when the decision is made a *queuing decision point*, shown in Fig. 4. It is apparent that we have a random time interval between the image capture and the queuing decision, which is at most a URB buffering period ($M \cdot U$). We call this *queuing decision jitter*.

We revisit an object's capture delay d_{capt} , defined as the time lapse between the physical appearance of an object and its successful capture. A successful capture means that the object is queued after its transmission. A capture can be unsuccessful when it is later dropped by the queue, because the queue does not have sufficient room to accommodate the frame associate with the object. It is apparent that d_{capt} is intertwined with the detector subsystem's service interval distribution \mathcal{S} . At the beginning of each object detection cycle, an image frame is retrieved from the queue, making room for a new image frame.

The queuing decision point now plays a key role in enqueueing the frame; the decision point must occur after this moment for the frame associated with it to be accepted by the queue.

In the best case, $d_{capt} = 0$ when an object appears right before a capture instant² and the queue has sufficient room to accommodate the image frame associated with the object. In contrast, the worst case occurs when an object appears right after the capture instant that makes the queue just become full so that all subsequently captured images should be dropped by the queue until the queue becomes available. The worst-case d_{capt} can now be drawn using the maximum number of image frames subsequently dropped within the longest object detection service interval³, denoted by s^{max} . We can estimate this number by counting the number of queuing decision points within s^{max} , by taking the effect of the maximum queuing decision jitter ($M \cdot U$) into account. The time interval determined by the number of queuing decision points can be shorter than the interval obtained by the same number of capture instants by at most $M \cdot U$. From this observation, s^{max} should be added by $M \cdot U$ to account for the maximum queuing decision jitter when it is divided by C , in order to obtain d_{capt} , which yields the range of d_{capt} , as follows:

$$0 \leq d_{capt} < \left\lceil \frac{s^{max} + M \cdot U}{C} \right\rceil C. \quad (3)$$

By summing Eq. (1) and Eq. (3), we have the minimum and maximum camera subsystem delays, denoted by d_{camera}^{min} and d_{camera}^{max} , respectively, as follows:

$$\begin{aligned} d_{camera}^{min} &= \left(\left\lceil \frac{I}{B} \right\rceil + 2 \right) U \text{ and} \\ d_{camera}^{max} &= \left\lceil \frac{s^{max} + M \cdot U}{C} \right\rceil C + \left(\left\lceil \frac{I}{B} \right\rceil + 2 + M \right) U. \end{aligned} \quad (4)$$

C. Detector Subsystem

The detector subsystem comprises three modules: (i) fetch, (ii) inference, and (iii) display. We now introduce these modules along with the delay components associated with them. The detector subsystem's service interval distribution \mathcal{S} will be identified after analyzing and profiling the three modules on an Nvidia Jetson AGX Xavier. For now, no other background workload is assumed.

(i) Fetch: The fetch module is responsible for retrieving an image frame from the queue. However, because V4L2 supports the streaming I/O method where only pointers to buffers are exchanged between the driver and applications [19], there is no actual memory copy for data transfer. Instead, most of its execution time is spent in resizing an input image to match the resolution of a DNN input layer. Therefore, the execution time of the fetch module, e_{fetch} , can be approximated as follows:

$$e_{fetch} \sim \mathcal{D}_{fetch}(r_{in}, r_{out}), \quad (5)$$

²In this paper, we do not consider the minimum exposure requirement of the camera. However, it can simply be assigned a value greater than zero as the minimum capture delay.

³During each object detection cycle, three sequential modules (fetch, inference, and display, which will be explained in the following section) are executed in a pipeline manner to detect the object. Hence, this interval is determined by the maximum among the delays incurred by these modules.

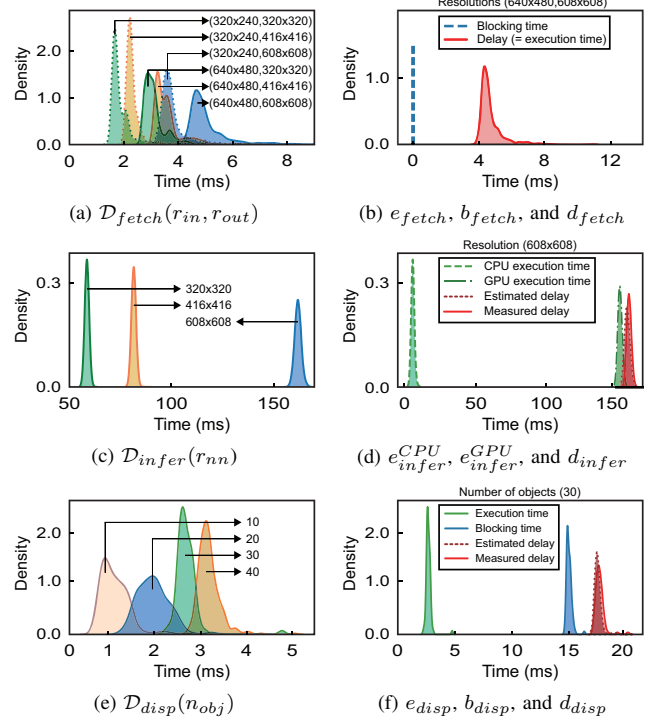


Fig. 6: Profiling results from YOLO v3.

where $\mathcal{D}_{fetch}(\cdot)$ is a probability distribution that is a function of input resolution (r_{in}) and output resolution (r_{out}) for resizing. Fig. 6(a) presents the profiling results of $\mathcal{D}_{fetch}(\cdot)$ with varying (r_{in}, r_{out}). Now, the delay of the fetch module, denoted by d_{fetch} , is defined as the time elapsed between its release (starting fetch thread) and its completion (ending image resize), which can be generalized as follows:

$$d_{fetch} = e_{fetch} + b_{fetch}, \quad (6)$$

where b_{fetch} is a blocking time factor. Note that, because no other background workload is assumed, no preemption delay is considered in d_{fetch} . Regarding b_{fetch} , a blocking system call in the fetch module may arise when an image is being retrieved from the queue. More specifically, when the fetch module tries to fetch an image while the queue happens to be empty, it is blocked until a new image arrives. Fig. 6(b) shows the profiling results of e_{fetch} , b_{fetch} , and d_{fetch} with fixed input and output resolutions presented above the plot. In the figure, $b_{fetch} = 0$ at all times, meaning that there were no empty queue cases. However, this depends on the system status, in particular, the queue status, which will be discussed in Section III-D.

(ii) Inference: The inference module performs forward propagation in a neural network, which is decomposed into CPU and GPU parts. In the CPU part, it simply calls asynchronous GPU operations with CUDA function calls, which include data copies between CPU and GPU memory as well as GPU kernel functions for the neural network. Then the GPU operations are queued into a CUDA stream, which is a FIFO (First In, First Out) queue maintained by the CUDA runtime.

The CPU thread blocks until the asynchronous GPU operations are completed by explicitly calling a CUDA synchronization function. Then the last GPU operation copies the detection results to the CPU memory. To run a single forward propagation, for example, YOLO v3 requires four copy operations and 322 GPU kernel operations (including repetitions). Note that the execution time of the inference module depends on the resolution of the input to the neural network r_{nn} , which is indicated by the profiling results in Fig. 6(c). Thus, the CPU execution time of the inference module e_{infer}^{CPU} and its GPU execution time e_{infer}^{GPU} are defined by random variables as follows:

$$\begin{aligned} e_{infer}^{CPU} &\sim \mathcal{D}_{infer}^{CPU}(r_{nn}) \text{ and} \\ e_{infer}^{GPU} &\sim \mathcal{D}_{infer}^{GPU}(r_{nn}), \end{aligned} \quad (7)$$

where $\mathcal{D}_{infer}^{CPU}(\cdot)$ and $\mathcal{D}_{infer}^{GPU}(\cdot)$ are probability distributions for CPU and GPU execution times, respectively, both of which are dependent on r_{nn} . For notational simplicity, the (unified) execution time of the inference module, e_{infer} , is defined as $e_{infer} = e_{infer}^{CPU} + e_{infer}^{GPU}$. Because there is no additional blocking factor, the delay of the inference module, d_{infer} , is given by

$$d_{infer} = e_{infer} \sim \mathcal{D}_{infer}(r_{nn}), \quad (8)$$

where $\mathcal{D}_{infer}(\cdot)$ can be estimated by the convolution of $\mathcal{D}_{infer}^{CPU}(\cdot)$ and $\mathcal{D}_{infer}^{GPU}(\cdot)$, as follows:

$$\mathcal{D}_{infer}(\cdot) = \mathcal{D}_{infer}^{CPU}(\cdot) \otimes \mathcal{D}_{infer}^{GPU}(\cdot). \quad (9)$$

Fig. 6(d) shows the profiling results with a fixed r_{nn} , where the estimated delay by Eq. (9) closely matches the measured delay. As shown in the figure, most of the execution time is spent in the GPU part.

(iii) Display: The display module visualizes the detection results by showing images with bounding boxes around each detected object. To accomplish this, the display module takes the following steps: first, the raw detection results with various confidence levels are filtered by a given confidence threshold. Next, irrelevant duplicates are joined (or removed) by the non-maximal suppression algorithm. Further, this module draws rectangles around each detected object. Thus, the execution time of the display stage, denoted by e_{disp} , is dependent on the number of detected objects n_{obj} , and is defined as

$$e_{disp} \sim \mathcal{D}_{disp}(n_{obj}). \quad (10)$$

In our measurements, we found that the number of raw detections and the final number of detected objects have a highly positive correlation. Thus, instead of using two variables, we use only the final number of detected objects when defining the probability distribution. Fig. 6(e) depicts the execution time distributions of the display stage against different values of n_{obj} . In addition, a blocking factor b_{disp} exists, which is caused by calling the OpenCV function that renders images on a display. Thus, the delay of the display module, d_{disp} , can be estimated by

$$d_{disp} = e_{disp} + b_{disp}. \quad (11)$$

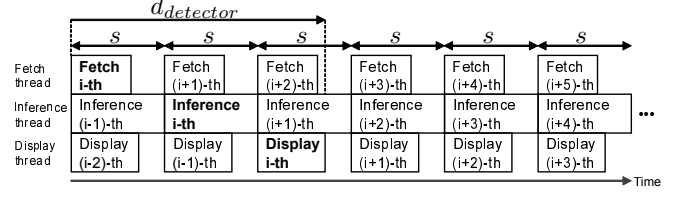


Fig. 7: Multithreaded pipeline architecture.

In the profiling results in Fig. 6(f), b_{disp} is significant because it involves the GUI subsystem of the operating system (e.g., X-window), which processes related pending events on the display window. The figure specifically shows the profiling results for $n_{obj} = 30$. For the worst-case analysis, we assume $n_{obj} = 30$ as the maximum number of detectable objects. Note that Darknet has a configurable limit in its configuration file. Moreover, even commercial object detection systems impose a limit on the possible number of detectable objects in their specifications. Note that although the display stage above explained is not always necessary in production vehicles, we consider it as a representative example of indispensable post-processing works.

Multithreaded pipeline architecture. In Darknet, the above three modules are organized with a multithreaded pipeline architecture to exploit multicore processors, using the fork-join model [22]. There are three parallel threads, one for each module: (i) fetch thread, (ii) inference thread, and (iii) display thread. At the beginning of an object detection cycle, the three threads are forked and then joined at the end of the loop to begin the next cycle. Fig. 7 shows the timeline of the multithreaded pipeline architecture. In general, during the i -th object detection cycle, the i -th image frame is fetched while the $(i-1)$ -th image frame is processed by the neural network, and the detection result of the $(i-2)$ -th image is displayed on the screen. In the figure, since the inference is the longest, the fetch and display threads should wait for the inference thread to complete at the end of each object detection cycle. Thus, the object detection *service interval*, s , is now generalized as

$$s = \max(\{d_{fetch}, d_{infer}, d_{disp}\}), \quad (12)$$

which is also referred to as object detection *cycle time* depending on the context. Regarding s , according to the profiling results in Fig. 6, d_{infer} is significantly longer than d_{fetch} and d_{disp} . Thus, in our experimental platform, we can simply assume a service interval distribution \mathcal{S} as

$$\mathcal{S} = \mathcal{D}_{infer}(\cdot). \quad (13)$$

However, in general, the service interval distribution \mathcal{S} of a detector subsystem by Eq. (12) relies on possible combinations of the values of three random variables $\{d_{fetch}, d_{infer}, d_{disp}\}$; the distribution can be defined as the set of the maximum among these three values for each of all possible combinations. Assuming $d_{fetch} = A$, $d_{infer} = B$, and $d_{disp} = C$, the corresponding probability is the product of the three respective probabilities, i.e., $\Pr\{d_{fetch} = A\} \times \Pr\{d_{infer} = B\} \times \Pr\{d_{disp} = C\}$. The final \mathcal{S} can be obtained by summing all

the probabilities for a given service interval. Based on \mathcal{S} , the delay of the detector subsystem, denoted by $d_{detector}$, can be represented as follows:

$$d_{detector} = 2s + d_{disp}, \quad (14)$$

signifying two object detection cycles plus one delay of the display thread, as depicted in Fig. 7. Then the minimum and maximum detector subsystem delays can be calculated as

$$\begin{aligned} d_{detector}^{min} &= 2s^{min} + d_{disp}^{min} \text{ and} \\ d_{detector}^{max} &= 2s^{max} + d_{disp}^{max}. \end{aligned} \quad (15)$$

D. Queue Delay

Based on the image arrival interval distribution \mathcal{A} and the object detection service interval distribution \mathcal{S} discussed in Section III-B and Section III-C, we can analyze the queue delay d_{queue} , which is defined as the time each image frame spends inside the queue. Darknet uses the OpenCV VideoCapture library for queue management with its default queue size $Q=4$. Recall that a full queue cannot accommodate additional incoming image frames. For the analysis of d_{queue} , we have to consider the following three cases:

Case 1: $\min(\mathcal{A}) > \max(\mathcal{S})$. In this case, object detection service intervals are always shorter (faster) than image arrival intervals. Then, from the perspective of the fetch thread, every time it attempts to retrieve an image frame, the queue is empty, thereby the fetch thread is blocked until an image frame arrives. Hence,

$$d_{queue} = 0. \quad (16)$$

In this case, b_{fetch} is eventually lengthened until $d_{fetch} = b_{fetch} + e_{fetch} = a$, where a is the image arrival interval. As a result, object detection cycle times are regulated by the image arrival intervals, regardless of d_{infer} and d_{disp} .

Case 2: $\max(\mathcal{A}) < \min(\mathcal{S})$. In this case, image arrival intervals are always shorter (faster) than object detection service intervals, such that whenever the fetch thread tries to retrieve an image frame, the queue is full. Thus, $b_{fetch}=0$ at all times. For the queue delay, after an image enters the tail of the queue, it requires Q object detection cycles (i.e., fetches) until it is dequeued. Thus, it takes roughly $Q \cdot s^{max}$ in the worst case. To be precise, the maximum queue delay is calculated as follows:

$$d_{queue}^{max} = Q \cdot s^{max} - (d_{tran}^{min} - M \cdot U). \quad (17)$$

The subtracted part reflects the minimum time that elapses between the beginning of an object detection cycle (i.e., making room to take newly captured images) and the actual arrival of an image frame. Note that the maximum queuing decision jitter ($M \cdot U$) should be considered because, at most, that amount of an image frame can be transmitted, after its capture, even before the object detection cycle begins. In the best case, the value of d_{queue}^{min} is calculated as follows:

$$d_{queue}^{min} = Q \cdot s^{min} - (d_{tran}^{max} + C). \quad (18)$$

This happens when an object appears just past the capture instant and has been successfully taken at the next capture

instant. Then the delay of up to C is accounted for in the capture delay, instead of the queue delay. In this case, we assume a zero queuing decision jitter to calculate the minimum queue delay.

Case 3: all other conditions. In this case, the queue status is non-deterministic. Its long-term average can be estimated only when the averages of the two distributions are equal. Moreover, because the queue length is limited within $[0, Q]$, the queue status at a certain instant cannot be estimated. Thus, the queue delay can be either $d_{queue}^{min} = 0$, as in **Case 1**, in the best case, or can be as much as Eq. (17), as in **Case 2**, in the worst case.

IV. END-TO-END DELAY OPTIMIZATION

Based on the analysis thus far, this section explains our approach to minimize the end-to-end delay in our object detection system. More specifically, the following three techniques will be explained: (i) on-demand capture, (ii) zero-slack pipeline, and (iii) contention-free pipeline. Fig. 8 shows how these techniques incrementally reduce the end-to-end delay from the original (vanilla) architecture to our final architecture. Note that the effect of our optimization methods depends on the relative execution times of the three threads. For ease of explanation, we assume that d_{infer} is significantly longer than d_{fetch} and d_{disp} , as illustrated in Fig. 8(a). In this case, the object detection cycle time s is dominated by d_{infer} , which is the most common case in embedded systems. The general applicability of our approach will be discussed in Section V.

A. On-demand Capture

From Eqs. (17) and (18) in Section III-D, it is apparent that we can reduce the queue delay by decreasing the queue size Q , especially for **Case 2**, and partly for **Case 3**. Note that even for **Case 1**, reducing Q has no negative side effect because it already works as though no queue exists. Motivated by the above observation, we can completely remove the queue, such that we only receive image frames when they are needed. We call this the *on-demand capture* method. For the implementation, we modified the OpenCV VideoCapture library. Listing 1 shows the pseudo code of the original OpenCV implementation where, in the initialization phase, a fixed queue size Q is requested from the driver with the ioctl command (REQBUFS) (Line 1). Then Q empty buffers, i.e., image holders, are put into the queue using the ioctl command (QBUF) (Line 2-3). After that, the fetch loop begins. Initially, it checks for the existence of an available image frame by making a select system call (Line 6). Upon returning from the select system call, it dequeues a buffer containing an image using the ioctl command (DQBUF) (Line 7), and then queues an empty buffer using the ioctl command (QBUF) (Line 8) again. It begins the next loop instantly. For comparison, Listing 2 shows our modified implementation with $Q=1$ (Line 1), which inserts an empty buffer into the queue using the ioctl command (QBUF) (Line 6) just before making the select system call. After the select system call returns, notifying an image arrival (Line 7), the buffer with the image frame is

dequeued by the `ioctl` command (`DQBUF`) (Line 8), and the next loop begins.

<pre> 1 ioctl(fd, REQBUFS, Q); 2 for(i=0; i<Q; i++) 3 ioctl(fd, QBUF, buf); 4 /* Image fetch loop */ 5 while(1){ 6 select(fd); 7 ioctl(fd, DQBUF, buf); 8 ioctl(fd, QBUF, buf); 9 } </pre>	<pre> 1 ioctl(fd, REQBUFS, 1); 2 3 /* Image fetch loop */ 4 while(1){ 5 select(fd); 6 ioctl(fd, QBUF, buf); 7 select(fd); 8 ioctl(fd, DQBUF, buf); 9 } </pre>
--	---

Listing 1: Original.

Listing 2: On-demand capture.

Fig. 8(a) illustrates the vanilla architecture with a queue, whereas Fig. 8(b) depicts our on-demand capture architecture. By comparing the two, it can be seen that our on-demand capture architecture no longer has a queue delay. Instead, it experiences a blocking time b_{fetch} when it makes the `select` system call, as in Listing 2. More specifically, the minimum and maximum blocking times are determined as follows:

$$\begin{aligned} b_{fetch}^{min} &= d_{tran}^{min} - M \cdot U \text{ and} \\ b_{fetch}^{max} &= d_{tran}^{max} + C. \end{aligned} \quad (19)$$

Note that they are the same as the subtracted parts of Eq. (17) and Eq. (18), respectively, which correspond to waiting times from the beginning of an object detection cycle until a new image frame, which will be used later, arrives at the queue, even though there is no blocking. In contrast, in the on-demand capture architecture, the fetch thread is blocked for the exact same amount of time as the waiting time, and once unblocked, the arriving image frame is fed directly into the fetch thread with no queue delay.

In Fig. 8(b), by the on-demand capture method described above, the queue delay is completely removed as

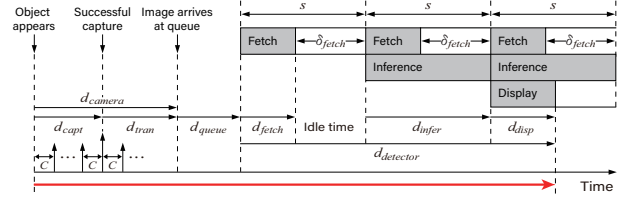
$$d_{queue} = 0. \quad (20)$$

As a side effect, in the figure, note that d_{camera} and d_{fetch} are somewhat overlapped by b_{fetch} , because the fetch thread is blocked waiting for an image arrival for b_{fetch} at its beginning. Thus, when calculating the detector subsystem delay, the overlapped amount of time (b_{fetch}), which is already accounted for in d_{camera} , should be subtracted from the original delay equations in Eq. (15), so we have the following update to the detector subsystem delay:

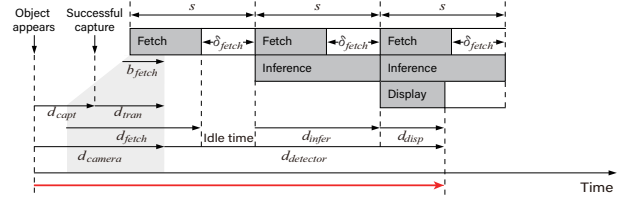
$$\begin{aligned} d_{detector}^{min} &= 2s^{min} + d_{disp}^{min} - b_{fetch}^{max} \text{ and} \\ d_{detector}^{max} &= 2s^{max} + d_{disp}^{max} - b_{fetch}^{min}. \end{aligned} \quad (21)$$

B. Zero-slack Pipeline

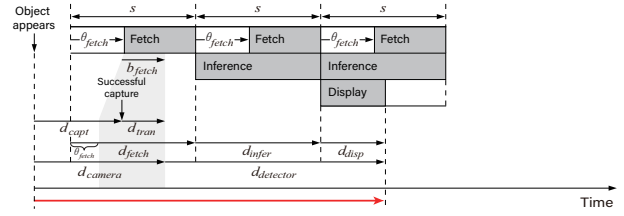
From Fig. 8(b), we can observe significant idle time gaps between the end of fetches and the beginning of the next object detection cycles, caused by the different lengths of the three pipeline stages, i.e., d_{fetch} , d_{infer} , and d_{disp} . If we can remove these idle times, the end-to-end delay can be further reduced. To that end, we introduce the notions of the slack and offset of the fetch thread, as follows:



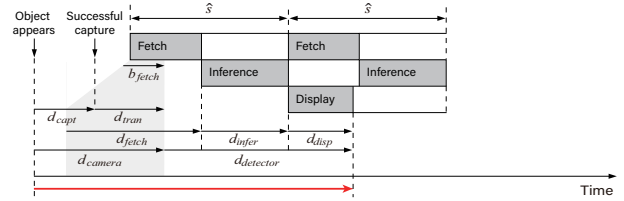
(a) Vanilla (original) architecture



(b) + On-demand capture: queue delays are removed.



(c) + Zero-slack pipeline: idle time gaps are removed.



(d) + Contention-free pipeline: memory contention is removed.

Fig. 8: End-to-end delay optimization methods. The thick red arrows below the time axes depict the end-to-end delays.

- **Fetch-thread slack (δ_{fetch}):** time gaps between the end of a fetch thread and the beginning of the next object detection cycle.
- **Fetch-thread offset (θ_{fetch}):** an artificial constant delay at every release of the fetch thread.

By accounting for an offset θ_{fetch} , d_{fetch} is increased to

$$d_{fetch} = \theta_{fetch} + e_{fetch} + b_{fetch}. \quad (22)$$

Then its slack can be calculated by

$$\delta_{fetch} = s - d_{fetch}, \quad (23)$$

where s denotes the object detection cycle time. Thus, we can increase θ_{fetch} until δ_{fetch} reaches zero (zero-slack), which eventually eliminates idle time gaps, as illustrated in Fig. 8(c). Then the detector subsystem delay is further reduced from Eq. (21) to

$$\begin{aligned} d_{detector}^{min} &= 2s^{min} + d_{disp}^{min} - b_{fetch}^{max} - \theta_{fetch} \text{ and} \\ d_{detector}^{max} &= 2s^{max} + d_{disp}^{max} - b_{fetch}^{min} - \theta_{fetch}. \end{aligned} \quad (24)$$

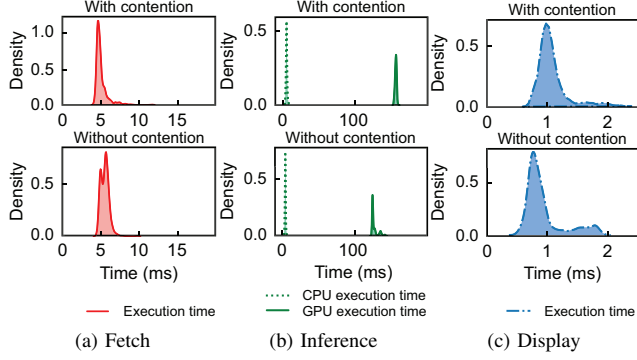


Fig. 9: Comparison of execution times with and without shared memory bandwidth contention.

However, applying too much θ_{fetch} may increase the object detection cycle time s itself. Thus, we must find an optimal θ_{fetch} that minimizes the end-to-end delay as much as possible while not increasing s . For that, θ_{fetch} is conservatively decided as

$$\theta_{fetch} = s^{min} - e_{fetch}^{max} - b_{fetch}^{max}, \quad (25)$$

such that the increased d_{fetch} barely touches the minimum object detection cycle time s^{min} (i.e., $\max(d_{fetch}) \approx s^{min}$). Due to the conservative decision, we cannot completely remove all idle time gaps. However, it is apparent that zero-slack pipeline with an appropriate θ_{fetch} reduces the end-to-end delay to some extent.

C. Contention-free Pipeline

When multiple threads execute on a shared memory multicore processor, there is significant memory bandwidth contention between concurrent threads [23], [24]. Moreover, when CPU threads run with GPU kernels on an integrated GPU, the effect is even worse [25]. Because Darknet uses a multi-threaded pipeline architecture, our method is not free from this. To quantify the memory contention effect, Fig. 9 shows the execution time distributions with and without contention. The execution times with contention are measured in the vanilla architecture, whereas the execution times without contention are measured while sequentially executing the object detection stages. In particular, Fig. 9(b) shows a significant increase in execution time (about 28 ms) for the GPU executions on the inference thread. In contrast, the fetch and display threads in Figs. 9(a) and 9(c) show relatively minor effects. Thus, it is clear that the inference thread suffers the most from the memory bandwidth contention. Based on this observation, our design choice is to isolate the inference thread to minimize the contention, whereas the fetch and display threads run concurrently as in the original architecture. Fig. 8(d) shows the resulting contention-free pipeline architecture.

Unlike other proposed methods described so far, which reduce the end-to-end delay without any trade-off, the contention-free pipeline method is not free. As shown in

Fig. 8(d), the object detection cycle time is changed (and possibly increased) to \hat{s} , which is defined as follows:

$$\hat{s} = \max(\{d_{fetch}, d_{disp}\}) + d_{infer}. \quad (26)$$

However, because d_{infer} is also decreased by contention-free executions, the increase from s to \hat{s} is marginal, which will be discussed in the experiments section. Then the detector subsystem delay is finally reduced as follows:

$$\begin{aligned} d_{detector}^{min} &= \hat{s}^{min} + d_{disp}^{min} - b_{fetch}^{max} \text{ and} \\ d_{detector}^{max} &= \hat{s}^{max} + d_{disp}^{max} - b_{fetch}^{min}. \end{aligned} \quad (27)$$

D. Considering Other Real-Time Workloads

Thus far, we have assumed no other workload in the system. This section discusses how our final pipeline architecture can be integrated with other real-time workloads. The most straightforward approach to ensure this integration is to isolate CPU cores required to parallelize each pipeline stage and a GPU, such that they are dedicated to our object detection system while the remaining CPU cores are shared by other workloads. However, this approach may still cause the same memory contention problem. To address this issue, we can use the RT-Gang scheduling framework [24], which eliminates inter-task interference by its one-gang-at-a-time policy. Here, a gang is a pre-defined group of tasks that is allowed to run in parallel. For example, in Fig. 8(d), we can define two gangs: (i) the fetch and display gang and (ii) the inference gang. Other real-time workloads can be grouped into a number of extra gangs, where all the gangs are scheduled by uniprocessor scheduling algorithms based on the RT-Gang framework.

Also, recall that the object detection cycle time \hat{s} in Eq. (26) is not a design parameter but the one that can be derived; we, therefore, should turn it into a design parameter (i.e., the period of our gangs) to use it in real-time scheduling. Then, if all the gangs are schedulable, the given object detection period will be guaranteed. Regarding the end-to-end delay, the following constraints should be met to benefit from its minimization: (i) within a gang period, the fetch and display gang should be completed prior to the beginning of the inference gang; (ii) the display thread should not be preempted by other real-time workloads. We set the fetch and display gang as the task with the highest priority in the system so as to satisfy the above constraints; this modification results in minimal end-to-end delay. The actual implementation and evaluation with the RT-Gang framework is left as our future work.

V. GENERAL APPLICABILITY OF OUR METHODS

This section discusses how our approach can be applied in more general settings in the following aspects:

To other object detection frameworks. Although we use Darknet throughout this study, our analysis and optimization methods regarding the camera subsystem and the queue delay can be applied to other object detection frameworks without modification. Further, because most other frameworks are designed without considering multicore processors, their detector subsystem can be enhanced by adapting our approach along

TABLE II: Default system configurations.

Parameter	Value
Camera frame rate (F)	30 fps
Camera cycle time (C)	33.333... ms
Camera resolution ($X \times Y$)	640×480
Bits per pixel (P)	16 (YUYV pixel format)
Queue size (Q)	4
YOLO neural network resolution	608×608

with Darknet’s multithreaded pipeline architecture. Thus, we consider that our solution can be generally used for the industry when building their production object detection systems, regardless of the framework they use.

To other DNNs. This study uses the YOLO DNNs because we are targeting object detection systems. However, as we do not exploit any DNN-specific features of YOLO, our methods can even be applied to other perception tasks (e.g., lane detection) with different DNN architectures. Nevertheless, we do not claim that our final architecture, as shown in Fig. 8(d), is generally optimal across all DNNs. For example, let us replace the YOLO inference stage with a lightweight lane detection DNN, which makes the pipeline stages well balanced, i.e., $d_{fetch} \approx d_{infer} \approx d_{disp}$. Then we cannot expect a meaningful delay reduction with our zero-slack pipeline. Thus, each optimization technique should be selectively applied after investigating the execution patterns of the target DNN.

To other hardware platforms. We conduct experiments on an Nvidia Jetson AGX Xavier, having an integrated GPU with significant memory contention between the CPU and the GPU. Although our analysis framework generally works for any GPU-based hardware platform, certain optimization techniques may not always provide as much benefit as in our experimental platform. For example, on other hardware platforms with a dedicated GPU with its isolated memory subsystem, our contention-free method will not provide satisfactory delay reductions as in our experimental platform. However, note that Nvidia Jetson AGX Xavier is a typical example of embedded systems for developing computer vision applications.

VI. EXPERIMENTS

A. Implementation

Our Real-Time Object Detector (R-TOD⁴) is implemented on an Nvidia Jetson AGX Xavier with 16 GB RAM, an 8-core ARM CPU, and a 512-core integrated Volta GPU. We use the Logitech C930e USB camera in the USB 2.0 high-speed mode. Among many, we selected the 640×480 resolution, 30 fps frame rate, and YUYV 16-bpp pixel format as our default camera configuration. As our software platform, we use Nvidia Ubuntu Linux-18.04 with JetPack-4.2.2, which is an SDK (Software Development Kit) for AI and computer vision applications, and a custom compiled OpenCV-3.3.1 library. We use the UVC camera driver in Linux. With regard to the application side, as the baseline object detector, the source code available at <https://github.com/AlexeyAB/darknet> is used.

⁴The source code is available at <https://github.com/aveeslab/r-tod>.

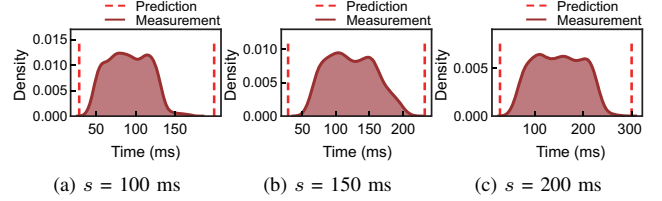


Fig. 10: Camera subsystem delays with varying service intervals (predictions and measurements).

Note that this is the most well-known fork of the original YOLO repository. In recent times, it has been regarded as the official repository because the original author of YOLO stopped his research in computer vision. Furthermore, for neural networks, we use YOLO v2, v3, and v4 models pre-trained using the MS COCO dataset, which are available in the same GitHub repository. Finally, we use 608×608 as the default input size of the neural network.

(i) On-demand capture. The OpenCV VideoCapture library is modified to make the queue size a tunable parameter. Because there is no such API in the library, we added an environment variable `QLEN`. If `QLEN=0`, on-demand capture is activated.

(ii) Zero-slack pipeline. We put a sleep function call at the beginning of the fetch thread. The amount of sleep θ_{fetch} is decided offline, and can be given by a command-line option when invoking the object detector.

(iii) Contention-free pipeline. We no longer fork the three threads when beginning an object detection cycle. Instead, only the fetch and display threads are forked. After the two threads join, the inference thread is started alone in isolation. The next object detection cycle begins instantly after the completion of the inference thread.

B. Evaluation of Our Analysis Methods

This section evaluates our end-to-end delay analysis framework presented in Section III. For the evaluation, YOLO v3 is used. However, note that we obtained similar evaluation results using YOLO v2 and v4 as well. Table II summarizes the default evaluation configuration, unless otherwise mentioned.

Fig. 10 shows the camera subsystem delay distributions where a camera subsystem delay is split into two parts: (i) d_{capt} and (ii) d_{tran} . Between the two, d_{tran} is actually measured, and d_{capt} is randomly generated to simulate an object’s random appearance between the previous and the current successful camera capture instants. Because the capture delay is tightly intertwined with the service interval of the detector subsystem, we measure the camera subsystem delays with varying service intervals. A dummy object detector with controllable artificial delays is used to control the service intervals. The figure shows that the camera subsystem delay distributions are within our predicted intervals as given by Eq. (4). Further, as reported in our analysis, the camera subsystem delay increases, from (a) to (c) in Fig. 10, with longer object detection service intervals.

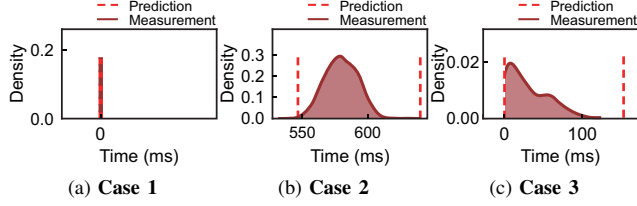


Fig. 11: Queue delays with the three cases (predictions and measurements).

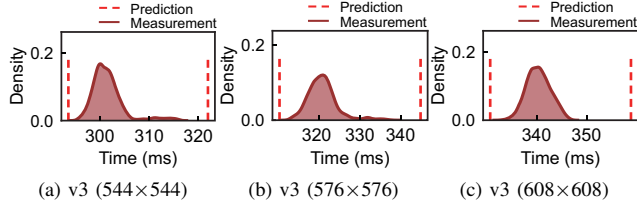


Fig. 12: Detector subsystem delays with varying neural network input resolutions (predictions and measurements).

Fig. 11 shows the measurement results of the queue delays and our predictions for the three cases considered when analyzing the queue delay. In Fig. 11(a) (**Case 1**), $d_{queue} = 0$ at all times as predicted in Eq. (16). In Fig. 11(b) (**Case 2**), the delay is predicted to be between 546 ms and 643 ms by Eq. (17) and Eq. (18) that successfully estimate the actual measurements. In Fig. 11(c) (**Case 3**), the queue status is non-deterministic as reported in our analysis. The measurement results are irregularly distributed between 0 ms and 103 ms. Note that even for **Case 3**, we can bound the minimum and maximum delays by reusing the analysis method of **Case 1** in the best case and the analysis method of **Case 2** in the worst case, respectively.

Fig. 12 presents the measurement results of the detector subsystem delay along with our predictions obtained using Eq. (15). The figures show that our predictions well estimate the detector subsystem delays with varying neural network resolutions. For the predictions, it is assumed that $n_{obj} = 30$ for the worst-case analysis. During the measurement, however, the actual number of detected objects is mostly less than 15. This is one of the reasons for the unavoidable pessimism in the worst-case prediction.

C. Evaluation of Our Optimization Methods

This section evaluates our end-to-end delay optimization methods presented in Section IV. We use the same system configuration as that shown in Table II, unless otherwise mentioned.

Fig. 13 shows the end-to-end delays and the object detection cycle times with varying queue sizes for YOLO v2, v3, and v4. In Fig. 13(a), end-to-end delays decrease linearly upon shrinking the queue size, where queue size of zero implies on-demand capture. Further, Fig. 13(b) shows that queue sizes do not affect the object detection cycle time. Thus, we can conclude that the queue between the camera subsystem and

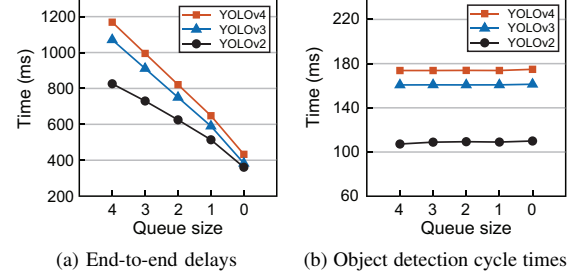


Fig. 13: Average end-to-end delays and object detection cycle times with varying queue sizes.

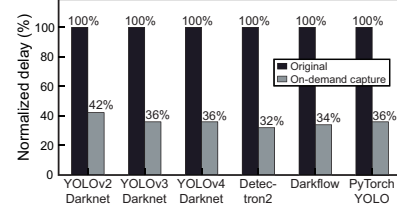


Fig. 14: Comparison of normalized average end-to-end delays before and after applying on-demand capture for various object detectors including PyTorch-based YOLO [26].

the detector subsystem is unnecessary, especially when trying to minimize the end-to-end delay.

Since many state-of-the-art object detectors commonly use the OpenCV VideoCapture library with $Q=4$, we tried to replace the original OpenCV library with our modified version anticipating that it could transparently reduce the end-to-end delay of arbitrary object detectors. We did not need to modify the object detectors, because OpenCV is a shared library. Fig. 14 shows how much the end-to-end delays of six different object detectors are reduced by our modified OpenCV library. This figure verifies that on-demand capture is generally applicable to other object detectors.

Fig. 15 shows our evaluation results for the optimization methods. We measured the end-to-end delays and object detection cycle times for YOLO v2, v3, and v4 by incrementally applying the three techniques. For the average system behavior, Figs. 15(a) and 15(b) show the average end-to-end delays and object detection cycle times, respectively. They reveal that the end-to-end delay is continually reduced from the vanilla architecture to the final architecture with the three optimization methods. The average delay reductions are 73% (v2), 76% (v3), and 76% (v4). Regarding the object detection cycle times, on-demand capture and a zero-slack pipeline do not increase the object detection cycle time, whereas a contention-free pipeline slightly increases the object detection cycle time. However, note that the increases are rather marginal. For the tail behavior, Figs. 15(c) and 15(d) show the 99th percentile end-to-end delays and object detection cycle times, respectively: the 99th percentile delays reduced by 60% (v2), 67% (v3), and 69% (v4).

As Darknet has been modified to improve the timing performance, we evaluate its impact on the *detection performance* with the KITTI MoSeg dataset [27], [28]. Be-

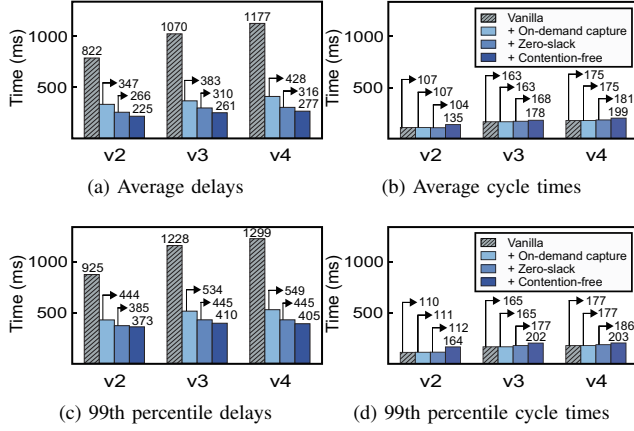


Fig. 15: Evaluation results (timing performance).

cause we cannot use a physical camera with datasets, we implemented a virtual USB camera by creating a V4L2 loopback device [29] that reads image files at a constant speed (i.e., 30 fps) and feeds them into the V4L2 queue such that Darknet cannot even notice it is virtual. The popular *mean average precision (mAP)* [17], [30] is used as the detection accuracy metric. After running 1000 real-world driving images, Fig. 16 indicates that our approach does not incur any penalty on the detection accuracy. The minute discrepancies are caused by different image capture timings due to the different object detection cycle times. In addition, the trend indicates that more recent YOLO versions show better results.

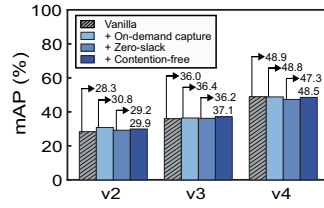


Fig. 16: Detection performance.

VII. RELATED WORK

Object detection algorithms. Recent object detection algorithms are based on DNNs and can be divided into two main categories. One category includes two-stage detectors such as R-CNN [31], Fast R-CNN [32], Faster R-CNN [33], and Mask R-CNN [34] that use a region proposal network to generate regions of interest in the first stage, where each region of interest is classified in the second stage. The other category includes single-stage detectors such as YOLO, SSD (Single Shot MultiBox Detector) [35], and RetinaNet [36] that perform region proposal and classification at the same time. In general, single-stage networks are faster but have slightly lower detection accuracy. Among them, YOLO is widely used in autonomous driving [5], [14], [15]. Although this study uses YOLO for the implementation and evaluation, our methods are generally applicable to any of above mentioned object detection algorithms.

Object detection frameworks. To train and deploy object detection DNNs, general-purpose DNN frameworks such as TensorFlow [37], Caffe [38], Caffe2 [39], and PyTorch [40]

can be used. In addition, various specialized object detection frameworks such as Darknet, Darkflow, Detectron [41], and Detectron2 are available. Among them, we use Darknet as our baseline implementation because it is written purely in C and CUDA, making it portable across different hardware platforms.

GPU resource management. To make efficient DNN-based real-time systems, it is important to understand the internal GPU scheduling. However, because this information is not disclosed to the public, many studies [42]–[47] have tried to devise formal GPU scheduling models by empirical methods. To minimize the shared memory bandwidth contention between CPU and GPU, BWLOCK++ [25] protects GPU kernels by performing memory bandwidth throttling, which can be used to augment our method. In [48], a parallel-pipeline execution model for Darknet is presented to improve throughput (i.e., number of cameras) with minor latency increases. On the contrary, our objective is to minimize the end-to-end delay assuming a single camera.

Real-time DNN inference. To support DNN inference with real-time requirements, S³DNN [49] provides system-level data fusion and supervised streaming for the efficient utilization of GPUs. A multi-path neural network [50] supports dynamic deadlines by managing multiple neural network execution paths with different execution times. In addition, Subflow [51] supports dynamic deadlines with selective sub-graph executions by taking varying time budgets into account. DART [52] supports the deterministic execution of real-time tasks while providing increased throughput to best-effort tasks. The aforementioned works can be used in conjunction with our method to guarantee explicit real-time requirements in multi-task environments.

VIII. CONCLUSION

This study was motivated by our observation that many state-of-the-art real-time object detectors show incomprehensibly large end-to-end time lags, despite having decent frame rates. In this light, we provide a complete understanding of the end-to-end delay of an object detection system, in particular, by using Darknet YOLO. On the basis of thorough investigation, we developed an end-to-end delay analysis framework that can precisely predict the best- and worst-case end-to-end delays. Furthermore, based on the analysis, we presented three optimization methods: (i) on-demand capture, (ii) zero-slack pipeline, and (iii) contention-free pipeline. By applying the above methods, we can reduce the average end-to-end delay of YOLO v3 by 76% and the 99th percentile delay by 67% on an Nvidia Jetson AGX Xavier.

In the future, we plan to extend our work with a self-optimizing pipeline framework that automatically reorganizes its pipeline architecture on varying neural networks and hardware platforms, possibly with more complex application scenarios, such as a sensor fusion system with multiple sensors.

ACKNOWLEDGMENT

This work was supported partly by the Korea Evaluation Institute Of Industrial Technology (KEIT) grant funded by the Ministry of Trade, Industry and Energy (MOTIE) (20000316, Scene Understanding and Threat Assessment based on Deep Learning for Automatic Emergency Steering), partly by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2014-3-00065, Resilient Cyber-Physical Systems Research), partly by the Ministry of Land, Infrastructure, and Transport (MOLIT), Korea, through the Transportation Logistics Development Program (20TLRP-B147674-03, Development of Operation Technology for V2X Truck Platooning), and partially by NSF grant CCF-1704859. J.-C. Kim is the corresponding author of this paper (Co-correspondence: K. Kang).

REFERENCES

- [1] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing deep neural network with pruning, trained quantization and Huffman coding," in *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: <http://arxiv.org/abs/1510.00149>
- [2] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [3] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, 2017, pp. 1–12.
- [4] Nvidia, "NVIDIA open source project," <https://github.com/nvidia/>, 2019.
- [5] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving: Constraints and acceleration," in *Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2018, pp. 751–766.
- [6] J. Redmon, "Darknet: Open source neural networks in C," <http://pjreddie.com/darknet/>, 2013–2016.
- [7] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.
- [8] —, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [9] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [10] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," <https://github.com/facebookresearch/detectron2>, 2019.
- [11] T. H. Trieu, "Darkflow," <https://github.com/thtrieu/darkflow>, 2018.
- [12] A. Navarro, R. Asenjo, S. Tabik, and C. Cascaval, "Analytical modeling of pipeline parallelism," in *Proceedings of the 18th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2009, pp. 281–290.
- [13] M. I. Gordon, W. Thies, and S. Amarasinghe, "Exploiting coarse-grained task, data, and pipeline parallelism in stream programs," in *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2006, pp. 151–162.
- [14] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monroy, T. Ando, Y. Fujii, and T. Azumi, "Autoware on board: Enabling autonomous vehicles with embedded systems," in *Proceedings of the 9th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPs)*, 2018, pp. 287–296.
- [15] M. Alcon, H. Tabani, L. Kosmidis, E. Mezzetti, J. Abella, and F. J. Cazorla, "Timing of autonomous driving software: Problem analysis and prospects for future solutions," in *Proceedings of the 26th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2020, pp. 267–280.
- [16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [17] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *International Journal of Computer Vision (IJCV)*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
- [18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.
- [19] The kernel development community, "Video for Linux API," <https://www.kernel.org/doc/html/v4.9/media/uapi/v4l/v4l2.html>, 2016.
- [20] G. Bradski, "The OpenCV Library," *Dr. Dobbs Journal of Software Tools*, 2000.
- [21] I. Lee, J. Y. Leung, and S. H. Son, *Handbook of Real-Time and Embedded Systems*. CRC Press, 2007.
- [22] M. E. Conway, "A multiprocessor system design," in *Proceedings of the Fall Joint Computer Conference*, 1963, pp. 139–146.
- [23] H. Yun, G. Yao, R. Pellizzoni, M. Caccamo, and L. Sha, "MemGuard: Memory bandwidth reservation system for efficient performance isolation in multi-core platforms," in *Proceedings of the 19th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2013, pp. 55–64.
- [24] W. Ali and H. Yun, "RT-Gang: Real-time gang scheduling framework for safety-critical systems," in *Proceedings of the 25th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2019, pp. 143–155.
- [25] —, "Protecting real-time GPU applications on integrated CPU-GPU SoC platforms," in *Proceedings of the 30th Euromicro Conference on Real-Time Systems (ECRTS)*, 2018, pp. 19:1–19:22.
- [26] A. Kathuria, "A PyTorch implementation of a YOLO v3 object detector," <https://github.com/ayoshkathuria/pytorch-yolo-v3>, 2018.
- [27] M. Siam, H. Mahgoub, M. Zahrn, S. Yogamani, M. Jagersand, and A. El-Sallab, "MODNet: Moving object detection network with motion and appearance for autonomous driving," *arXiv preprint arXiv:1709.04821*, 2017.
- [28] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [29] Umlaute, "v4l2loopback - a kernel module to create V4L2 loopback devices," <https://github.com/umlaute/v4l2loopback>, 2010–2020.
- [30] S. Belousov, "mAP: Mean average precision for object detection," https://github.com/bes-dev/mean_average_precision, 2020.
- [31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.
- [32] R. Girshick, "Fast R-CNN," in *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
- [33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.
- [34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the 16th IEEE international conference on computer vision (ICCV)*, 2017, pp. 2961–2969.
- [35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.
- [36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the 15th IEEE international conference on computer vision (ICCV)*, 2017, pp. 2980–2988.
- [37] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
- [38] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia (MM)*, 2014, pp. 675–678.

- [39] Facebook Open Source, “Caffe2: A new lightweight, modular, and scalable deep learning framework,” <https://caffe2.ai>, 2017.
- [40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.
- [41] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, “Detectron,” <https://github.com/facebookresearch/detectron>, 2018.
- [42] N. Otterness, V. Miller, M. Yang, J. Anderson, F. D. Smith, and S. Wang, “GPU sharing for image processing in embedded real-time systems,” in *Proceedings of the 12th Annual Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPERT)*, 2016, pp. 23–29.
- [43] N. Otterness, M. Yang, T. Amert, J. Anderson, and F. D. Smith, “Inferring the scheduling policies of an embedded CUDA GPU,” in *Proceedings of the 13th Annual Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPERT)*, 2017, pp. 47–52.
- [44] N. Otterness, M. Yang, S. Rust, E. Park, J. H. Anderson, F. D. Smith, A. Berg, and S. Wang, “An evaluation of the Nvidia TX1 for supporting real-time computer-vision workloads,” in *Proceedings of the 23rd IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2017, pp. 353–364.
- [45] T. Amert, N. Otterness, M. Yang, J. H. Anderson, and F. D. Smith, “GPU scheduling on the Nvidia TX2: Hidden details revealed,” in *Proceedings of the 38th IEEE Real-Time Systems Symposium (RTSS)*, 2017, pp. 104–115.
- [46] M. Yang, N. Otterness, T. Amert, J. Bakita, J. H. Anderson, and F. D. Smith, “Avoiding pitfalls when using Nvidia GPUs for real-time tasks in autonomous systems,” in *Proceedings of the 30th Euromicro Conference on Real-Time Systems (ECRTS)*, 2018, pp. 20:1–20:21.
- [47] S. Jain, I. Baek, S. Wang, and R. Rajkumar, “Fractional GPUs: Software-based compute and memory bandwidth reservation for GPUs,” in *Proceedings of the 25th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2019, pp. 29–41.
- [48] M. Yang, S. Wang, J. Bakita, T. Vu, F. D. Smith, J. H. Anderson, and J.-M. Frahm, “Re-thinking CNN frameworks for time-sensitive autonomous-driving applications: Addressing an industrial challenge,” in *Proceedings of the 25th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2019, pp. 305–317.
- [49] H. Zhou, S. Bateni, and C. Liu, “S³DNN: Supervised streaming and scheduling for GPU-accelerated real-time DNN workloads,” in *Proceedings of the 24th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2018, pp. 190–201.
- [50] S. Heo, S. Cho, Y. Kim, and H. Kim, “Real-time object detection system with multi-path neural networks,” in *Proceedings of the 26th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2020, pp. 174–187.
- [51] S. Lee and S. Nirjon, “SubFlow: A dynamic induced-subgraph strategy toward real-time DNN inference and training,” in *Proceedings of the 26th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2020, pp. 15–29.
- [52] Y. Xiang and H. Kim, “Pipelined data-parallel CPU/GPU scheduling for multi-DNN real-time inference,” in *Proceedings of the 40th IEEE Real-Time Systems Symposium (RTSS)*, 2019, pp. 392–405.