3-2022

MRIM: Enabling Mixed-Resolution Imaging for low-power pervasive vision tasks

Jiyan WU
Singapore Management University, jiyanwu@smu.edu.sg

Vithurson SUBASHARAN
Singapore Management University

Tuan TRAN
Singapore Management University, tuantran@smu.edu.sg

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

MRIM: Enabling Mixed-Resolution Imaging for Low-Power Pervasive Vision Tasks

Ji-Yan Wu, Vithurson Subasharan, Tuan Tran and Archan Misra
Singapore Management University.
Email: {jiyanwu, vithursons, tuantran, archanm}@smu.edu.sg.

***Abstract*—While many pervasive computing applications increasingly utilize real-time context extracted from a vision sensing infrastructure, the high energy overhead of DNN-based vision sensing pipelines remains a challenge for sustainable in-the-wild deployment. One common approach to reducing such energy overheads is the capture and transmission of lower-resolution images to an edge node (where the DNN inferencing task is executed), but this results in an accuracy-vs-energy tradeoff, as the DNN inference accuracy typically degrades with a drop in resolution. In this work, we introduce *MRIM*, a simple but effective framework to tackle this tradeoff. Under *MRIM*, the vision sensor platform first executes a lightweight preprocessing step to determine the saliency of different sub-regions within a single captured image frame, and then performs a saliency-aware non-uniform downscaling of individual sub-regions to produce a "mixed-resolution" image. We describe two novel low-complexity algorithms that the sensor platform can use to quickly compute suitable resolution choices for different regions under different energy/accuracy constraints. Experimental studies, involving object detection tasks evaluated traces from two benchmark urban monitoring datasets as well as a prototype Raspberry Pi-based *MRIM* implementation, demonstrate *MRIM's* efficacy: even with unoptimized embedded platform, *MRIM* can provide system energy savings of 35+% or increase task accuracy by 8+%, over conventional baselines of uniform resolution downscaling or image encoding, while supporting high throughput.**

***Index Terms*—Mixed resolution, imaging tasks, energy consumption.**

I. Introduction

Vision-based sensing, typically using a network of infrastructurally-deployed cameras, is an important enabler for a variety of pervasive computing applications, such as situation awareness [3], human activity detection [13], shopper behavior analytics [2] and vehicular traffic monitoring [25]. Such expanded use of vision-based sensing has been accelerated by the reduced cost of high-resolution vision sensors and the impressive accuracy gains achieved by DNNs for tasks such as object detection [16] and object recognition [28]. The high energy overhead of visual sensing pipelines, however, continues to remain a major obstacle to its more widespread adoption, especially for in-the-wild deployments in spaces such as forests, parks and highways.

Given the increasing importance of developing ultra-low power or a battery-less sensing infrastructure [8], a variety of approaches have explored the development of low-power vision sensing systems. Most such low-power vision systems adopt an offloading-based architecture, where the pervasive sensor platform simply *captures* and *wirelessly transmits* (possibly preceded by some lightweight encoding) to a more-resourced (e.g., GPU-equipped) edge node, where the actual DNN-based AI pipelines are executed. Even so, pervasive applications continue to suffer from the *fidelity-vs.-energy* tradeoff: reduction in energy consumption is achieved by sacrificing either resolution (spatial granularity) or frame rate (temporal granularity), which in turn affects the DNN inference accuracy.

In this work, we explore the use of a novel information-centric approach, ***M**ixed-**R**esolution **IM**aging* (**MRIM**), as a means of improving this fidelity-vs.-energy tradeoff. The *MRIM* approach (illustrated in Figure 1a) hypothesizes that the operational lifetime of the sensor platform could be increased if we could find a *lightweight* mechanism to reduce the volume of transferred data (and thus the dominant transmission energy cost) without affecting the subsequent DNN inference accuracy. Under this approach (as illustrated in Fig. 1b, where 4 different sub-regions are processed at two distinct resolutions), the individual images captured by a camera sensor are broken up into multiple sub-regions, with the different sub-regions then down sampled at different resolutions prior to transmission. Our proposed approach, involving differential resolution *within* a single frame, is distinct from prior work on dynamic camera resolution adaptation [9], [15], which assume that any single frame is acquired, processed and/or transmitted at a uniform spatial resolution.

Intuitively, MRIM enables a more judicious use of system-resources on an energy-constrained vision sensor platform, adopting a lower resolution budget for the low-priority areas, while conserving the usage of higher resolution for the region of interest. The MRIM approach is motivated by two intuitive observations: (a) in most event-monitoring scenarios, objects or activities of interest are often not spread out uniformly over a camera's entire field-of-view (FoV) but localized or concentrated in certain *salient* sub-regions of the captured image; and (b) because the resolution reduction (and the corresponding loss in information fidelity) is directed preferentially towards the regions with lower saliency, the overall accuracy of DNN-based vision tasks remains largely unaffected. Building a practical MRIM-based vision-based sensing approach is, however, a non-trivial task and must address the following key research questions:

- How can the camera platform determine the saliency of different sub-regions within an image (a necessary pre-requisite before applying the principle of mixed-resolution downsampling)? In particular, to ensure that any savings in transmission energy are not negated by a higher processing energy overhead, it is essential that this determination be computationally cheap and incur low latency.
- Can the relationship between the energy overheads vs. vision

(a) Mixed-Resolution Image Transfer (b) Example of Mixed-Resolution Image (c) Accuracy gain (MRIM vs. Baselines)

Fig. 1: MRIM Paradigm and Overall Benefit

task accuracy (for varying resolution values) be accurately estimated across a diverse set of image/environmental context (e.g., for both images with a large number of small objects or a small number of large objects)?

- Given such saliency and energy-vs.-accuracy estimates, how can the camera platform determine, in a computationally lightweight manner, the right levels of resolution reduction (fidelity reduction) to be applied to each such sub-region?

Through our work, we show that it is indeed feasible to develop a practical MRIM approach that can overcome these challenges, focusing specifically on the commonplace *object detection* task. In particular, we shall (a) propose and demonstrate the efficacy of mixed-resolution determination algorithms that significantly reduce transmission bandwidth/energy without adversely affecting the accuracy of state-of-the-art DNN-based object detectors, and (b) develop a lightweight, low-power and accurate saliency determination approach that executes on the camera platform. Using a combination of a real-world prototype and diverse, real-world image traces, we shall demonstrate that *MRIM* can provide either a total systems-level energy savings of $\approx 35\%$ or task accuracy improvement of 8% over current approaches of uniform resolution reduction or image encoding.

Key Contributions: We make the following key contributions:

- *Introduce the MRIM framework:* Through systematic studies, we establish both (a) the tradeoff between the visual resolution of objects and the resulting accuracy of DNN object detectors, and (b) the non-uniform spatial properties of such objects in the FoV of typical pervasive camera deployments. These insights help motivate the principle of mixed-resolution imaging, which preferentially preserves pixel resolution in sub-regions with a greater *predicted* number of objects, while degrading the resolution of less salient sub-regions.
- *Devise & Evaluate Mixed-Resolution Algorithms:* We formulate the MRIM problem as one of (i) either minimizing the total image transmission size subject to a mean task accuracy constraint, or conversely, (ii) maximizing the task accuracy subject to a maximum energy budget. We then present two novel algorithms: (a) *Max-Min*, which preferentially increases the resolution of higher saliency sub-regions until the image-level objectives are satisfied, and (b) *Water-Filling*, which incrementally increases the resolution of all individual sub-regions equitably, until the image-level objectives are achieved. Via experimental studies with two different tasks/datasets–(a) human detection using WildTrack [4] and (b) vehicle detection using CityScapes [5]—we show that our proposed algorithms can provide $\sim 10-20\%$ improvement in object detection accuracy compared to currently-adopted approaches of either image encoding or uniform resolution adjustment. Fig. 1c summarizes the performance gains achieved by our *MRIM* strategy.
- *Demonstrate the Overall Effectiveness of an MRIM-based System:* We build and evaluate a working prototype of an *MRIM*-based camera, using the RPi (v3) board. The prototype integrates the mixed-resolution algorithms with a lightweight object detection technique (empirically shown to incur only 10 mJ/frame energy overhead) to determine the saliency of different sub-regions. Through careful experimental studies, we demonstrate that, in spite of many non-ideal system characteristics (e.g., high baseline power consumption), our *MRIM* approach provides energy savings/frame of $33-36\%$ and 28%, respectively, over the uniform resolution and image encoding approaches, while achieving equivalent object detection accuracy. Overall, *MRIM* allows the operational lifetime of such pervasive vision sensors to be doubled without loss in task accuracy.

II. Related Work

MRIM draws upon prior work in both (a) adaptive resolution in image capture, and (b) energy/power optimization in intelligent vision sensing systems.

A. Low Power Camera/Vision Sensing

LiKamWa et al. [15] demonstrated that image sensor energy consumption is ideally proportional to frame rate and resolution, and suggested multiple techniques (clock frequency control and low-power standby mode) to further reduce sensing power. To reduce the energy overhead of vision-related tasks, prior approaches utilize either on-board image processing (e.g., Cyclops [22]), a combination of low & high resolution cameras (e.g., SensEye [12]) or selective event-triggered activation of power-hungry vision sensors (e.g., Glimpse [20]). Several novel approaches have developed ultra-low power or battery-less camera sensors—for example, WISPCam [19] uses an RFID-powered harvester to trigger the capture of low-resolution, low frame-rate images, [18] utilizes analog backscatter communication to transfer HD-quality video transfer from an energy-harvesting vision sensor, while Elf [27] supports object counting by solar-powered cameras by adaptively adjusting the frame rate. In almost all cases, these approaches either require specialized hardware or additional

infrastructure (e.g., RFID readers) and usually support low quality, infrequent ($<$ 1 FPS) image capture. Collaborative sensing, across multiple cameras, has also been used to reduce the per-sensor energy overheads by opportunistically *deactivating* selected cameras-e.g., EECS [6] uses knowledge of (a) each camera's object detection accuracy, and (b) potential energy overhead to select a preferred set of $\{$activated cameras, video processing parameters$\}$ to monitor a common region.

B. Mixed-Resolution Image Processing

Past work has studied the broad relationship between image resolution and accuracy of vision-based tasks. The Banner prototype [9] demonstrated how dynamically reducing the *overall* image resolution (in contrast to *MRIM's* approach of utilizing differential resolution within a single image), based on the object's distance, can reduce camera sensing energy by 70%. The concept of *image resizing*, as a means of accelerating the computation of vision tasks, has also been recently explored in [10], where different regions of a single image are reduced, at an edge node, to different sizes (based on their priority), thereby increasing the overall inferencing throughput. Vision-based systems have also explored the processing of mixed-resolution *multiview* videos, where different cameras capture images at different spatial resolution and content from higher-resolution video streams is used to upscale the images captured by lower-resolution cameras (e.g., [17], [24]).

C. Efficient On-board Image Processing

We shall see that *MRIM's* success lies partly in being able to determine the saliency of different image sub-regions in an ultra-lightweight manner. Light-weight neural detection detection models, such as Haar feature [1], LFFD [7] and libface [21]), have been proposed for on-board execution. To support accurate object detection, approaches such as MobiSR [14] utilize a cheaper, low-resolution camera for image capture, followed by on-board upscaling on mobile devices to generate super-resolution images.

III. Motivating the *MRIM* Approach

Our overall approach for low-power pervasive vision utilizes the system architecture illustrated Fig. 2. In this architecture, the vision sensor platform performs the following key functions: (a) image capture–i.e., using the sensor to capture the raw image; (b) image pre-processing–i.e., performing any functions (such as compression) locally prior to transmission; and (c) image transmission–i.e., using a suitable networking interface (e.g., WiFi/4G) to transfer the processed image to an edge/cloud device. The edge/cloud platform then performs the vision task by executing the DNN pipeline; to keep the pervasive sensor cost and energy overheads low, we assume that the sensor platform does not have specialized hardware (e.g., GPUs) and cannot thus support efficient, high-throughput execution of the complex state-of-the-art DNN models, such as YOLO v3 [23].

Our focus is purely on reducing the total power/energy consumption of the sensor platform (without compromising on the eventual accuracy of the vision task), such that this sensor platform can operate for longer duration without needing recharging. To achieve this, the sensor platform performs additional pre-processing via two conceptually distinct functional components: (a) **Saliency Estimator**: as a precursor to performing diffential resolution downscaling, it determines the saliency of different regions in the image frame by estimating the likely general location and other relevant attributes of objects of interest. Note that, for high frame rate video, saliency determination need not be performed on each frame, but only intermittently (e.g., once every 1-2 secs), as object attributes are unlikely to dramatically vary over O(msec) timescales; (b) **Resolution Adjuster**: this component, which lies at the hear of *MRIM*, modifies the resolution of each sub-region of the captured image, taking into account the region's saliency and the resulting accuracy-vs.-energy tradeoffs.

MRIM's requires careful consideration of the tradeoff between the Pre-processing and Transmission energy overheads: intuitively, the additional steps of saliency estimation and resolution adjustment will result in increased pre-processing energy, which should be offset by a greater reduction in transmission energy. In addition, we will need to show that our proposed approach offers a superior energy-vs.-accuracy profile compared to two established baselines for reducing transmission overheads: (i) Image compression/encoding, where standard codecs (often implemented in hardware) are used to perform lossy compression of the image/video content, and (ii) Uniform resolution adaption, where the entire image is uniformly downscaled (without consideration of the image content) to a specified size. Determining the right choices for *MRIM* thus first requires a careful understanding of both (i) the energy overheads of different pre-processing mechanisms and the subsequent transmission phase, and (ii) the resulting impact of different image sizes/resolutions on the DNN inference accuracy.

TABLE I: Mathematical notations

$\mathbb{N}_d$	no. of detections	$\mathbb{N}_r$	no. of sub-regions
mAP	target mean average precision	$\mathcal{R}_i\|_{1\leq i\leq N}$	image regions
$\mathcal{S}_i\|_{1\leq i\leq \mathbb{N}}$	confidence scores	$\mathbb{E}$	energy constraint
E	total energy consumption	S	image file size
$\mathcal{V}_i\|_{1\leq i\leq \mathbb{N}}$	regional resolution values	Est_Eng	est. energy consumption

A. Modeling System Energy Consumption

We first model the energy consumption for processing and transmitting target image frames captured by an embedded camera platform (Table I lists the basic mathematical notations used). Specifically, the image processing (with our resolution adjustment algorithm or JPEG encoding) and data transmission (using 3G/4G/Wi-Fi chip) account for the main portion of power consumption of image application.

The total system energy consumption E including the idle (baseline) energy, as well as the energy spent in image capture, processing and subsequent transmission, is represented as:

$$\mathsf{E} = \mathsf{E}_{\text{idle}} + \mathsf{E}_{\text{cap}} + \mathsf{E}_{\text{proc}} + \mathsf{E}_{\text{trans}}, \tag{1}$$

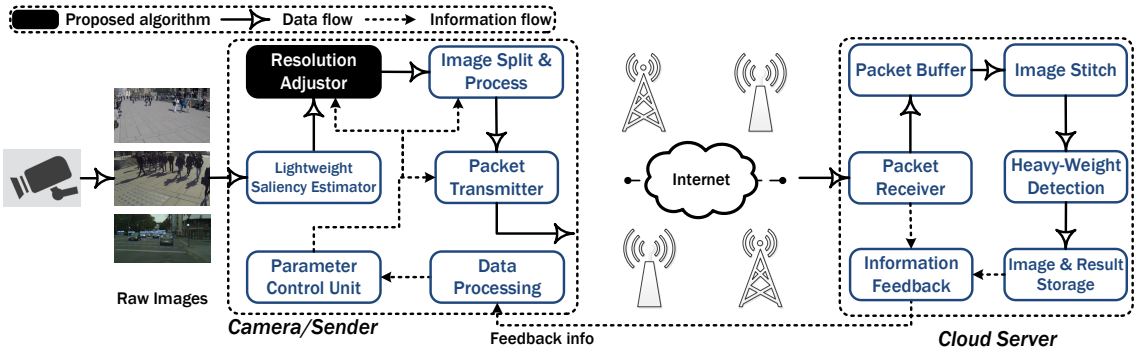


Fig. 2: System design of the proposed accuracy and energy aware mixed-resolution image transmission framework.

where $\mathsf{E}_{\text{idle}}, \mathsf{E}_{\text{cap}}, \mathsf{E}_{\text{proc}}, \mathsf{E}_{\text{detect}}$ represent the idle state (i.e., baseline), image capture, processing and transmission energy, respectively. The idle baseline energy includes energy spent in powering the different system components (e.g., processor, SD card, etc.). The transmission energy depends on the data transmission power P_{tran} and duration (i.e., the amount of data transferred), and can be represented as $\mathsf{E}_{\text{trans}} = \mathsf{P}_{\text{tran}} \cdot d$; also, the capture energy E_{cap} is usually negligibly smaller (O(μW), compared to O(mJ)) than the other components.

The processing energy incurred consists of both E_{sal}, the energy spent in saliency estimation (if this step is required), and $\mathsf{E}_{\text{image}}$, the subsequent energy spent in modifying the captured image–i.e.,:

$$\mathsf{E}_{\text{proc}} = \mathsf{E}_{\text{sal}} + \mathsf{E}_{\text{image}}. \tag{2}$$

For *MRIM*, $\mathsf{E}_{\text{image}}$ includes the energy spent in computing the modified resolution values and the subsequent downsampling; for a conventional compression-based approach, $\mathsf{E}_{\text{image}}$ would represent the energy spent in lossy compression.

Lightweight Saliency Estimator: *MRIM* relies on the use of a lightweight model to derive the approximate count/distribution of objects in each image sub-region and thereby derive each region's saliency. To ensure *MRIM's* overall energy efficiency, it is important to characterize the energy profile of such candidate lightweight models. As measured via an implementation on the Raspberry Pi 3B platform, a Haar feature based detector [26] is able to achieve as low as 200 mJ for each iteration, and is shown to provide adequate indication of the likely presence of objects in individual sub-regions. In practice, this estimator can be run intermittently (e.g., once every 1/2 seconds or 20/30 frames), resulting in a very low normalized energy overhead (∼1-2mJ/frame).

B. DNN Accuracy vs. Resolution

The *MRIM* approach is premised on the observation that the accuracy of DNN-based object detectors depends on the resolution (reflecting the information fidelity) of the underlying images. To understand this phenomenon in detail, we consider a typical state-of-the-art DNN model, such as YOLO v5 [11]. The output of such an object detector includes the class id (e.g., a person or vehicle object), bounding box coordinates (the center point, width and height) and the confidence score (a value between (0,1) that represents the probability of the bounding box containing an object. To represent accuracy, we adopt the widely-used mean Average Precision (mAP) metric, which computes the mean AP over all classes and/or overall IoU (Intersection over Union) thresholds as follows: $mAP = \frac{\sum_{k=1}^{N_c} AP_k}{N_c}$, where N_c is the number of classes and AP_k indicates the average precision for the k^{th} class.

(a) 1350*900 (b) 225*75

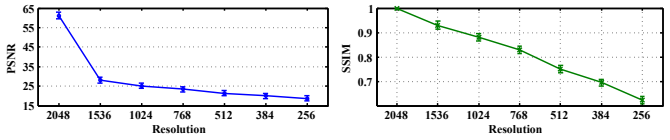
Fig. 3: DNN Detection Accuracy vs. Image Resolution.

As an illustration of our underlying hypothesis, Fig. 3 plots the bounding boxes and confidence values identified by the YOLO v5 object detector [11] on two images of the same scene, but at different levels of resolution (original= $1350*900$ in Fig. 3a, reduced= $255*75$ in Fig. 3b. We can clearly observe both a decrease in the number of detected vehicles, as well as a substantial reduction in the confidence scores of the detected objects. On closer inspection, we see that the 'smaller size" vehicles appearing in the upper half part of the image suffer a higher accuracy loss than those "larger-sized" vehicles in the lower half. Although the mAP is reduced from 69.3% to 56.8% due to the resolution downgrade, the file size S of the underlying image exhibits a 12-fold reduction, from 707 to 57 KB, which should (as per Equation 1) lead to a reduction in the total energy consumption. Motivated by these observations, we now conduct a deeper study of the relationship between the system energy E, deep learning model accuracy mAP and image quality.

C. Energy-Quality-Accuracy (EQA) Tradeoff

Generally speaking, the model accuracy is proportional to image *fidelity*, with the fidelity itself correlated to the resolution of the input image. To illustrate this point, we study how the fidelity/quality of images varies as the overall image resolution is progressively decreased. Fig. 4 plot the mean and confidence intervals (based on a corpus of 500 images curated from the WildTrack and CityScapes datasets) of two widely

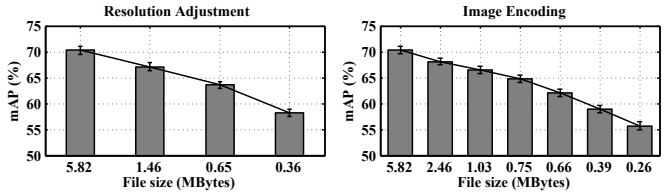
used measures of objective and subjective image quality, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity), respectively, as the input image resolution varies from 2048*2048 to 256*256. We see that a reduced resolution leads to a progressive loss in quality, due to the degradation in the color and positional features of the underlying objects.



(a) PSNR (b) SSIM

Fig. 4: Image quality vs. resolution.

Similar to image quality, the accuracy of DNN-based vision tasks should also increase with the underlying image resolution. However, after a certain point, any increase in image resolution provides only a marginal improvement in vision task accuracy. As mentioned earlier, a reduction in image resolution has two effects: it decreases the mAP of task accuracy as well as the size of the underlying image files. To capture this relationship, Fig. 5 plots the mAP vs. file sizes (resolution) for two different baseline strategies: (a) uniform resolution downsampling and (b) compressive image encoding (using the principle component analysis compression technique). The figure plots the mean and 95% confidence intervals, computed over the 500 representative images mentioned earlier. In addition, we also use our Raspberry Pi implementation (Section V) to empirically measured the resulting processing latency and energy consumption.



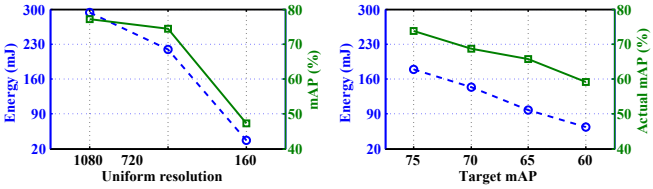
(a) resolution (b) encoding

Fig. 5: mAP versus file size.

We observe that both approaches, uniform downscaling and encoding, exhibit almost identical mAP-vs.-size tradeoff. In addition, the improvement in mAP values is more dramatic at low file sizes and becomes more muted as the image resolution increases from medium to high resolution–e.g., an ∼4-fold increase in image size from 1.46MB to 5.82MB results in an mAP increase of ≤2%. In addition, on carefully analyzing the mAP performance for individual images, we observe that:

- For images with predominantly larger-sized objects, the mAP degradation is not significant as the file size decreases. Conversely, the mAP values for images with predominantly small objects exhibit a much steeper drop as the file size (image resolution) decreases.
- While the mAP-vs.-file size variation is similar for both resolution downscaling and image encoding, the two strategies differ in their computational cost and latency. In particular, the image compression approach incurs much higher latency and energy (avg.=261 msecs and 73.8mJ) than the resolution downscaling approach (avg.=141 msecs and 53.7 mJ).

The EQA Tradeoff: Combining the individual experimental results allows us to now understand the energy-vs.-accuracy tradeoff generated by changing image quality (reflected by different image sizes). We conduct the experiments in two ways: (i) evaluate the energy and detection accuracy (precision) of different uniform resolutions (from 160p to 1080p) and (ii) later, gradually decrease the resolution values of the image regions of the 1080p image until the mAP reduces to {75%,70%,65%,60%}. Fig. 6 plots the results, for both the uniform downscaling approach and, for comparison, the differential techniques (*MRIM*) that we shall detail in Section IV.



(a) uniform resolution (b) mixed-resolution

Fig. 6: Energy versus mAP.

The results in Fig. 6a indicate the crux of the problem: under uniform resolution, *both* energy and precision drop significantly when resolution drops from 720p to 160p. However, as shown in Fig. 6b, permits a much more gentle decrease in mAP compared to the linear drop in energy. By deliberately and gradually changing the resolution of individual regions, we can achieve approximately the same level of precision (as shown in the green plot) with lower energy (blue plot) than that achieved via uniform resolution downscaling.

D. Problem Formulation

MRIM's choice of the downscaled resolutions for each sub-region can then be formulated as one of two distinct objective functions:

- P1: Given a minimum mean average precision, adjust the resolution of each image sub-region so as to minimize the total energy consumption, i.e.,

$$\mathsf{P1}: \mathcal{V}_i|_{1\leq i\leq \mathbb{N}} = \underset{mAP\geq\overline{mAP}}{\arg\min} \{\mathsf{E}\}.$$

  For this objective (suited for scenarios where the vision task has a minimum required fidelity), the accuracy serves as as a lower-bound constraint (e.g., $\overline{mAP} \geq 75\%$).
- **P2**: Given an energy constraint E, adjust the resolution values of each image sub-region so as to maximize the DNN task accuracy, i.e.,

$$\mathsf{P2}: \mathcal{V}_i|_{1\leq i\leq \mathbb{N}} = \underset{\mathsf{E}\leq\overline{\mathsf{E}}}{\arg\max} \{\text{mAP}\}.$$

  In this case (suited for scenarios where the platform has a finite battery capacity and a target lifetime), E serves as an upper-bound constraint (e.g., $\overline{\mathsf{E}} \leq 50$ mJ/frame).

Given the many real-world non-ideal characteristics in both system energy consumption and DNN performance, developing a provably optimal solution to each problem is infeasible and impractical. Hence, we shall next focus on developing efficient (low-complexity) heuristic algorithms for P1 and P2.

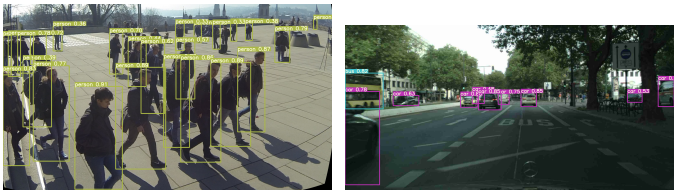
IV. Resolution Adjustment Algorithms

We now describe two different algorithms that the Resolution Adjuster can use to determine the different resolution choices for each of the sub-regions. The algorithm design is driven by our observation (using open-source image datasets corresponding to two representative tasks, human detection and vehicle detection, and illustrated in Fig. 7 below) that most captured images exhibit one of two spatial characteristics:

(i) *Uniform spatial distribution* (see Fig. 7a), where objects of interest are typically distributed across the entire image, even though the size of the objects vary depending on the observer-object distance.

(ii) *Skewed Distribution* (see Fig. 7b) , where objects of interest are typically observed in selected salient sub-regions of the image–e.g., mostly confined to the upper or left portion of the camera's FoV.

In addition, as observed earlier in Fig. 3, reduced resolution impacts the detection accuracy for smaller-sized objects disproportionately, implying that the algorithms must also incorporate the different resolution-to-mAP relationship for different *object sizes*. Given these observations, (i) the Max-Min algorithm tends to allocate higher resolution disproportionately to a smaller number of high saliency areas until the image satisfies a minimum predicted mAP value, and is thus better suited for images with skewed spatial distribution, whereas (ii) the Water-Filling algorithm, which conceptually attempts to equalize the predicted mAP values of all sub-regions, is better suited for images with uniform spatial distribution.



(a) Even distribution (b) Skewed distribution

Fig. 7: Even & Skewed Spatial Distributions

A. Estimating Resolution-to-mAP Values

To enable the *MRIM* algorithms to determine the right resolution choices, we first need to build a predictive estimate of how the final DNN task accuracy will be affected by different possible downscaled resolution candidates. To estimate this, we first compute the *weighted confidence* of the objects detected by the Saliency Estimator as: $WConf = \frac{\sum_{i=1}^{\mathbb{N}} S_i \cdot bbox_i}{\sum_{i=1}^{\mathbb{N}} bbox_i}$, where $bbox$ represents the bounding box area and S represents the class confidence of each detected object. Subsequently, we utilize a look-up table (which has been populated by extensive empirical studies) that maps this confidence value to an overall image mAP predicted to be achieved, for different image resolution/size values, when the heavyweight DNN (YOLOv5 [11] in our case) is executed on the edge node. For example, a sample entry in the lookup table might be of the form $(WConf = 57.2, Res = 300, mAP = 67.6)$, implying that that a Saliency Estimator WConf value of 57.2% is estimated to eventually result in a YOLO mAP=$\approx$ 67.6%, if the image was downsized to a 300x300 resolution.

Algorithm 1: Accuracy-Aware Energy Minimization

Input: $\mathbb{N}_d$, $\mathbb{N}_r$, bounding boxes, $\mathcal{S}_i|_{1\leq i\leq \mathbb{N}}$, $\overline{mAP}$, $\mathcal{V}_i|_{1\leq i\leq \mathbb{N}}$;
Output: $\mathcal{V}_i|_{1\leq i\leq N}$, $\mathcal{Q}_i|_{1\leq i\leq N}$, mAP;

Calculate mAP based on confidence scores for each class;
if $mAP < \overline{mAP}$ **then**
  Rank the regions $\mathcal{R}_i|_{1\leq i\leq \mathbb{N}_r}$ based on included bounding box areas in descending order;
  for *each region* $\mathcal{R}_i|_{1\leq i\leq \mathbb{N}_r}$ **do**
    for $\Delta\mathcal{V}$ *from small to large resolution changes* **do**
      $\mathcal{V}_i = \mathcal{V}_i + \Delta\mathcal{V}$;
      $\mathcal{S}_j = \mathcal{S}_j + \overline{\mathcal{S}}, \forall$ object $j \in \mathcal{R}_i$;
      $\mathcal{Q}_j = \mathcal{Q}_j - \Delta\mathcal{V}$;
      $mAP = \frac{\sum_{i=1}^{\mathbb{N}} \mathcal{S}_i}{\mathbb{N}}$;
      if $mAP \geq \overline{mAP}$ **then**
        break;
      end
    end
  end
end
else
  Rank the regions $\mathcal{R}_i|_{1\leq i\leq N}$ based on included bounding box areas in ascending order;
  for *each region* $\mathcal{R}_i|_{1\leq i\leq N}$ **do**
    for $\Delta\mathcal{V}$ *from small to large resolution changes* **do**
      $\mathcal{V}_i = \mathcal{V}_i - \Delta\mathcal{V}$;
      $\mathcal{S}_j = \mathcal{S}_j + \overline{\mathcal{S}}, \forall$ object $j \in \mathcal{R}_i$;
      $mAP = \frac{\sum_{i=1}^{\mathbb{N}} \mathcal{S}_i}{\mathbb{N}}$;
      if $mAP < \overline{mAP}$ **then**
        $\mathcal{V}_i = \mathcal{V}_i + \Delta\mathcal{V}$;
        break;
      end
    end
  end
end
return $\mathcal{V}_i|_{1\leq i\leq N}$, mAP;

B. Max-Min Algorithm

The Max-Min algorithms operate in greedy fashion, preferentially adjusting the resolution of individual sub-regions in order of descending priority, until the overall image mAP reaches the specified threshold. The priority order is determined by the saliency of individual sub-regions (which is precomputed based on the number of objects predicted by the

Lightweight Saliency Estimator) as well as an evaluation of whether the initial overall image mAP is lower or higher than the target mAP. At a high-level, if the current overall image AP is higher than the target mAP, we seek to preferentially increase the resolution of the regions with highest saliency; conversely, if the current overall image AP is higher than the target mAP, we preferentially decrease the resolution of the regions with the lowest saliency.

Algorithm 1 provides the high-level pseudocode of the Max-Min algorithm for objective P1. The algorithm inputs include the number of sub-regions $\mathbb{N}_r$, number of objects $\mathbb{N}_d$ and their bounding boxes (detected by the Lightweight Saliency Estimator), and target mean average precision $\overline{mAP}$. The algorithm starts with a nominal (low) resolution allocation to each of the sub-regions, and uses the afore-mentioned lookup table to estimate (line 1) each sub-region's anticipated mAP score, as well as the image's overall mAP. In case the estimated mAP is below the target mAP, the algorithm proceeds to progressively increase the resolution of individual sub-regions individually, starting with the most salient sub-region (the one with the largest number of predicted objects). At each step of such resolution adjustment, it recomputes the overall mAP (lines 6-9) and stops wheneve this overall mAP has exceeded the target value (lines 10-12). However, if the overall image mAP has not reached the target value even after the first sub-region's resolution has been maximally increased, the algorithm then greedily proceeds to the region with the next highest saliency. This process is repeated until the overall mAP target has been achieved or all possible regions have been expanded to the maximum permissible resolution. Conversely, if the initial estimated overall mAP is higher than the target mAP, the algorithm assumes that the current resolution choices are too generous and seeks to iteratively decrease the resolution of sub-regions, starting with the lowest saliency region, until it 'just' exceeds the target mAP. The complexity of Algorithm 1 is $O(\mathbb{N}_r \cdot \mathbb{N}_d \cdot \frac{V}{\Delta V})$. We shall show in Section V that, under reasonable values of $\mathbb{N}_r$ (=4, 6, 8,...), the complexity of this algorithm is low enough to permit low-latency, low-energy execution on embedded platforms.

The overall greedy approach of Max-Min can also be suitably adapted to tackle the optimization problem P2 with the energy constraint $\overline{\mathcal{E}}$, as detailed in Algorithm 2. As before, the algorithm operates in greedy fashion, prioritizing individual sub-regions on the basis of their estimated saliency. In this case, however, during each iteration of resolution adjustment, the total energy consumption for that specific region (and thus the overall image) is re-estimated (using Equation 1), with the adaptation process continuing until the estimated total energy is 'just' below the permitted energy budget.

C. Water-Filling Algorithm

The Water-Filling Algorithms are inspired by prior work on equalizing channel performance in communication systems, and are based on the observation that water height effectively equalizes across multiple connected reservoirs independent of their individual heights. At a high-level, the algorithm views the resolution (pixel count) as a fluid resource that is effectively distributed among the different sub-regions so as to equalize their water *height*, where the height is defined by each region's predicted mAP value.

Algorithm 2: Energy-Constrained Accuracy Maximization

Input: $\mathbb{N}_d$, $\mathbb{N}_r$, bounding boxes, $\mathcal{S}_i|_{1\leq i\leq \mathbb{N}}$, $\mathbb{E}$, $\mathcal{V}_i|_{1\leq i\leq \mathbb{N}}$;
Output: $\mathcal{V}_i|_{1\leq i\leq N}$, $\mathcal{R}_i|_{1\leq i\leq N}$, Est_Eng;

1 $\mathcal{R}_i = \mathcal{R}_i^{\min}, \forall 1 \leq i \leq N$;
2 Rank the regions $\mathcal{R}_i|_{1\leq i\leq N}$ based on included bounding box areas in descending order;
3 **for** *each region* $\mathcal{R}_i|_{1\leq i\leq N}$ **do**
4 **for** $\Delta\mathcal{V}$ *from small to large resolution changes* **do**
5 $\mathcal{V}_i = \mathcal{V}_i + \Delta\mathcal{V}$;
6 $\mathcal{S}_j = \mathcal{S}_j + \overline{\mathcal{S}}, \forall$ object $j \in \mathcal{R}_i$;
7 $\mathcal{Q}_j = \mathcal{Q}_{\min}$;
8 $E_i = E_i^{\text{idle}} + E_i^{\text{enc}} + E_i^{\text{tran}}$;
9 Est_Eng $= \sum_{i=1}^{N} E_i$;
10 **if** *Est_Eng* $\geq \overline{\mathbb{E}}$ **then**
11 $\mathcal{V}_i = \mathcal{V}_i - \Delta\mathcal{V}$;
12 break;
13 **end**
14 **end**
15 **end**
16 return $\mathcal{V}_i|_{1\leq i\leq N}$, Est_Eng;

Algorithm 3 provides the high-level pseudocode for this approach for objective P1 (minimizing energy under an accuracy constraint). The algorithm starts by assuming each sub-region to be associated with the lowest permitted spatial resolution (pixel count). The resolution level is then incrementally increased across the board, and the resulting average accuracy is computed. Subsequently, each sub-region's individual predicted mAP score (height) is compared against this image-wide average, and the resolution for that region is iteratively increased until it is no longer below the current image-wide average value. This iterative approach effectively causes regions with higher saliency (larger number of objects or smaller-sized objects) to benefit from an increased allocation of pixel account, thereby assuring that such regions do not suffer poor accuracy. The Water-Filling approach can also be similarly adapted to objective P2 (maximizing accuracy under an energy constraint), although the pseudocode is omitted due to space constraints.

V. Performance Evaluation

We evaluate the performance of our proposed algorithms, both in terms of its (a) the resulting object detection task accuracy (mAP), and (b) the actual energy and latency overheads on real-world embedded vision sensor platforms. Accuracy performance is evaluated by replacing images from multiple benchmark public datasets. For the energy and latency metrics, we evaluate *MRIM* using the *Raspberry Pi (RPi) 3B*. RPi 3B is a popular embedded platform, equipped with Imx219 image sensor, BCM2837, a quad-core 1.2GHz ARM-Cortex processor, 1GB of LPDDR2 SDRAM and an onboard BCM43438

Algorithm 3: Water-Filling Resolution Adjustment With Accuracy Requirement

Input: $\mathbb{N}_d$, $\mathbb{N}_r$, bounding boxes, $\mathcal{S}_i|_{1\leq i\leq \mathbb{N}}$, $\overline{\text{mAP}}$, $\mathcal{V}_i|_{1\leq i\leq \mathbb{N}}$;
Output: $\mathcal{V}_i|_{1\leq i\leq N}$, $\mathcal{R}_i|_{1\leq i\leq N}$, Est_Eng;

Cur_ACC $= \frac{\sum_{i=1}^{\mathbb{N}} \mathcal{S}_i}{\mathbb{N}}$;
if *Cur_ACC* $<$ *mAP* **then**
  for *each region* $\mathcal{R}_i$ **do**
    Cur_ACC$_{\mathcal{R}_i} = \frac{\sum_{j=1}^{\mathbb{N}} \mathcal{S}_{\mathcal{R}_i}}{\mathbb{N}_{\mathcal{R}_i}}$;
    while *Cur_ACC*$_{\mathcal{R}_i} < \overline{\text{mAP}}$ **do**
      $\mathcal{V}_i = \mathcal{V}_i + \Delta\mathcal{V}$;
      Cur_ACC $= \frac{\sum_{i=1}^{\mathbb{N}} \mathcal{S}_i}{\mathbb{N}}$;
      $E_i(\mathcal{V}_i) = E_i^{\text{idle}}(\mathcal{V}_i) + E_i^{\text{tran}}(\mathcal{V}_i)$;
    end
  end
end
Est_Eng $= \sum_{i=1}^{N} E_i(\mathcal{V}_i)$;
return $\mathcal{V}_i|_{1\leq i\leq N}$, Est_Eng;

chip supporting an 802.11ac radio. Power measurements, both system-level and for individual functional components, are obtained via the use of the Monsoon power monitor.

A. Alternative Baselines

We compare *MRIM's* differential downscaling approach (via either Max-Min or Water-filling algorithms) with baselines of:

- *Compressive Image Coding:* We utilize the JPEG encoder of Python 3.8, and modify the compression quality values (higher value $\rightarrow$ higher quality) within the range $(100, \ldots, 20)$ to execute different levels of compression.
- *Uniform Resolution Downscaling:* In this approach, the entire image is uniformly downscaled to the target resolution, without considering the saliency of different sub-regions.

We compare with image encoding method because this is commonly used in image transmission and storage systems.

B. Datasets

We utilize two different public benchmark datasets that are representative of typical pervasive vision applications:

- *WildTrack:* The WildTrack dataset [4] involves the use of multiple HD 1920x1080 cameras (with an average $\sim$55% FoV overlap across cameras) to capture a very crowded public area on a university campus for the purposes of *human object detection*.
- *CityScapes:* The CityScapes dataset [5] includes 5000 images, across 27 cities, representing a wide variety of urban environments, and annotated with objects corresponding to 30 classes (e.g., vehicle, construction, human). In our studies, we utilize a subset of 2000 images most suited for the *vehicle object detection* task.

C. Energy-Accuracy Tradeoff

We first study how the total energy of the vision platform, as well as the task accuracy (YOLOv5 based object detection) is affected by different resolution choices, expressed by file size constraints. Fig. 8 plots the average energy consumption vs. mAP, across both WildTrack and CityScapes, as measured using the RPi platform. In addition, Figure 1c (presented earlier in Section I) demonstrates the tradeoff between mAP and transmission file size for the different approaches. Overall, we see that both Max-Min and Water-filling outperform the other baselines–in particular, the *MRIM* approaches are able to achieve $\approx$20% increase in accuracy under identical energy overheads. We also note that, as expected, the mAP of Uniform Downscaling degrades rapidly as the image is downscaled. While Image Encoding can achieve, on average, higher mAP, its energy overhead is significantly higher compared to all other approaches.

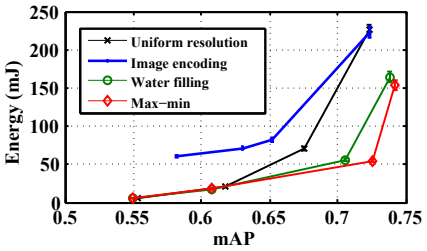


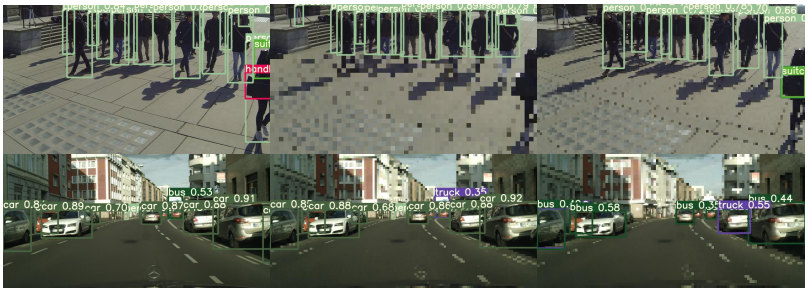
Fig. 8: Energy vs. Accuracy: *MRIM* vs. Alternatives

D. Max-Min vs. Water-filling

To further study the differences between the two algorithms presented in Section IV, Table II plots the energy-vs.-mAP variation for two different scenarios: (i) Water-Filling applied to objective **P1** and (ii) Max-Min applied to objective **P2**, when the raw camera image has a resolution of 1350*900. We can see that both approaches are able to dramatically reduce the overall file size (by $\approx$ 80-65%, compared to an original image size of $\sim$ 1.2 MB) with only a modest 3.6% loss of accuracy (to $\sim$ 0.68 from the maximum value of 0.73, achieved when the raw image is input to the DNN). *MRIM* can thus achieve at least a significant reduction in system energy, in spite of the RPI's non-negligible baseline power of 230 mW. Also, Fig. 9 illustrates the resulting compressed images (and the object detection output) under both schemes, for one representative image from each dataset. We can see that Max-Min is more aggressive in reducing the resolution for less salient regions (notice the increased blockiness of the foreground areas for WildTrack), whereas Water-filling tends to preserve greater detail for such lower-saliency regions.

E. Delay & Energy Characteristics:

To further illustrate the appeal of *MRIM's* computationally-efficient, yet effective, technique of image downsampling, Table III plots the variation in file size/energy consumption and processing latency (on the RPi) as the degree of compression is varied. We see that while compressive encoding can indeed reduce the file size significantly, the additional on-board processing overhead dampens the reduction in overall latency and energy, relative to *MRIM*. For example, under compressive encoding, a file size of 96KB incurs a processing latency of



(a) Original Images (b) Max-Min Algorithm (c) Water-Filling Algorithm

Fig. 9: Original vs. *MRIM*-based Images

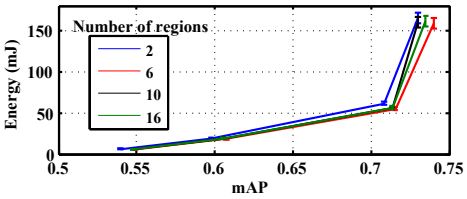


Fig. 10: Energy vs. Accuracy (Varying $\mathbb{N}_r$)

TABLE II: Water-filling vs. Max-Min Performance.

Water filling (accuracy-aware)			
Index	**File size (KB)**	**Total (mJ)**	**Target mAP**
1	472.8	107.52	0.673
2	119.5	27.62	0.642
3	52.09	10.83	0.568
4	17.6	3.472	0.526
Max-min (energy constraint)			
Index	**File size (KB)**	**Target energy (mJ)**	**mAP**
1	715.3	156.878	0.687
2	183.5	43.836	0.663
3	69.07	13.93	0.593
4	18.79	3.854	0.539

141 msecs; in contrast, with *MRIM*, a reduced file size of 69 KB incurs a processing latency of only 3.5 msecs.

TABLE III: *MRIM*vs. Image Encoding.

Resolution Adjustment			
Resolution	**File size (KB)**	**Total (mJ)**	**latency (ms)**
1350*900	736	156.9	196
775*450	219	43.8	50.8
338*225	69.07	13.9	14.3
225*75	19	3.9	3.56
Image encoding			
Jpeg quality	**File size (KB)**	**Total (mJ)**	**latency (ms)**
100	736	156.9	196
80	143.4	54.4	145
60	96.43	46.2	141
20	51.87	39.6	134

VI. Discussion

While our work attests to the power of *MRIM*, there are a few additional open issues requiring further investigation.

Choice of $\mathbb{N}_r$: We experimentally observe that the most suitable values for the number of distinct sub-regions (an input to our algorithms) are in the range of $[4, 6, 8, 10, 12]$. Figure 10 plots the energy vs. mAP variation for different values of $\mathbb{N}_r$. Intuitively, a larger value of $\mathbb{N}_r$ permits more fine-grained resolution adjustment (and thus improved mAP); however, the increased algorithm complexity leads to a larger value of E_{proc}. Indeed, we can see that a choice of $\mathbb{N}_r = 6$ provides higher accuracy at lower energy overheads, compared to $\mathbb{N}_r = 10$ or 12 . As the optimal value of $\mathbb{N}_r$ depends on the deployment-specific spatial properties of objects, we plan to develop a technique that allows a specific sensor instance to autonomously determine its optimal choice of $\mathbb{N}_r$.

Extension to Other Vision Tasks: While we considered only the object detection task, we believe that the general phenomenon of "reduced DNN accuracy with reduced image resolution" is more general, implying that *MRIM* should be applicable for a wider variety of vision tasks (e.g., scene classification). However, the accuracy-vs.-resolution tradeoff can be discontinuous for different tasks–e.g., it is likely that person identification requires a minimum resolution to operate.

Low Power Vision Platforms: We chose the RPi platform for our experiments primarily for expediency (easily available, open source drivers, etc.). However, the RPi's baseline/idle power of 230mW is significantly higher than other specialized ultra-low power platforms (such as Pixy2[1]), making the system far from energy-proportional [15] and thus limiting the gains exhibited by *MRIM*. We believe that *MRIM* can achieve substantially higher energy reduction ($\sim$70-80%) when applied to such specialized sensing platforms.

VII. Conclusion

MRIM introduces an approach for performing differential resolution downscaling on different regions of a single image, so as to reduce the overall energy overheads of pervasive vision sensing without compromising the accuracy of DNN-based vision tasks. We have described simple, but effective and computationally efficient, techniques for both dynamically estimating the saliency of different sub-regions of a single captured image, and subsequently determining the resolution values for each sub-region. Via the use of multiple benchmark urban monitoring datasets and an RPi-based implementation, we have demonstrated that *MRIM* can consume $\sim$30% less energy than even a hardware-optimized image encoder and achieve $\sim$20% improvement in object detection accuracy over a comparable uniform resolution downscaling approach.

Acknowledgement

This work was supported by the National Research Foundation, Singapore under its NRF Investigatorship grant (NRF-NRFI05-2019-0007). 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.

[1]CMUCam5 (Pixy-2 platform) URL- http://www.cmucam.org/projects/cmucam5

REFERENCES

[1] A. Bahadur. Haar feature-based light-weight detector. https://github.com/akshaybahadur21/FaceDetection, 2017.

[2] C. Bermejo, D. Chatzopoulos, and P. Hui. Eyeshopper: Estimating shoppers' gaze using cctv cameras. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2765–2774, New York, NY, USA, 2020. Association for Computing Machinery.

[3] N. Bicocchi, M. Lasagni, and F. Zambonelli. Bridging vision and commonsense for multimodal situation recognition in pervasive systems. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 48–56. IEEE, 2012.

[4] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[6] T. Dao, K. Khalil, A. K. Roy-Chowdhury, S. V. Krishnamurthy, and L. Kaplan. Energy efficient object detection in camera sensor networks. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 1208–1218, 2017.

[7] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan. Lffd: A light and fast face detector for edge devices. *arXiv preprint arXiv:1904.10633*, 2019.

[8] J. Hester and J. Sorber. The future of sensing is batteryless, intermittent, and awesome. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, SenSys '17, New York, NY, USA, 2017. Association for Computing Machinery.

[9] J. Hu, A. Shearer, S. Rajagopalan, and R. LiKamWa. Banner: An image sensor reconfiguration framework for seamless resolution-based tradeoffs. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 236–248, 2019.

[10] Y. Hu, S. Liu, T. Abdelzaher, M. Wigness, and P. David. On exploring image resizing for optimizing criticality-based machine perception. In *2021 IEEE 27th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pages 169–178, 2021.

[11] G. Jocher. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. https://github.com/ultralytics/yolov5, Oct. 2020.

[12] P. Kulkarni, D. Ganesan, P. Shenoy, and Q. Lu. Senseye: A multi-tier camera sensor network. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05. ACM, 2005.

[13] T. Kumrai, J. Korpela, T. Maekawa, Y. Yu, and R. Kanai. Human activity recognition with deep reinforcement learning using the camera of a mobile robot. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2020.

[14] R. Lee, S. I. Venieris, L. Dudziak, S. Bhattacharya, and N. D. Lane. Mobisr: Efficient on-device super-resolution through heterogeneous mobile processors. In *The 25th Annual International Conference on Mobile Computing and Networking*, MobiCom '19. Association for Computing Machinery, 2019.

[15] R. LiKamWa, B. Priyantha, M. Philipose, L. Zhong, and P. Bahl. Energy characterization and optimization of image sensing toward continuous mobile vision. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 69–82, 2013.

[16] L. Liu, H. Li, and M. Gruteser. Edge assisted real-time object detection for mobile augmented reality. MobiCom '19, New York, NY, USA, 2019. Association for Computing Machinery.

[17] S.-P. Lu, S.-M. Li, R. Wang, G. Lafruit, M.-M. Cheng, and A. Munteanu. Low-rank constrained super-resolution for mixed-resolution multiview video. *IEEE Transactions on Image Processing*, 30:1072–1085, 2020.

[18] S. Naderiparizi, M. Hessar, V. Talla, S. Gollakota, and J. R. Smith. Towards battery-free hd video streaming. In *Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation*, NSDI'18, page 233–247. USENIX Association, 2018.

[19] S. Naderiparizi, A. N. Parks, Z. Kapetanovic, B. Ransford, and J. R. Smith. Wispcam: A battery-free rfid camera. *2015 IEEE International Conference on RFID (RFID)*, pages 166–173, 2015.

[20] S. Naderiparizi, P. Zhang, M. Philipose, B. Priyantha, J. Liu, and D. Ganesan. Glimpse: A programmable early-discard camera architecture for continuous mobile vision. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '17, page 292–305. Association for Computing Machinery, 2017.

[21] H. Peng and S. Yu. A systematic iou-related method: Beyond simplified regression for better localization. *IEEE Transactions on Image Processing*, 30:5032–5044, 2021.

[22] M. Rahimi, R. Baer, O. I. Iroezi, J. C. Garcia, J. Warrior, D. Estrin, and M. Srivastava. Cyclops: In situ image sensing and interpretation in wireless sensor networks. In *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, SenSys '05. ACM, 2005.

[23] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[24] T. Richter, J. Seiler, W. Schnurrer, and A. Kaup. Robust super-resolution for mixed-resolution multiview image plus depth data. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(5):814–828, 2015.

[25] A. Skordylis and N. Trigoni. Efficient data propagation in traffic-monitoring vehicular networks. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):680–694, 2011.

[26] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, 2001.

[27] M. Xu, X. Zhang, Y. Liu, G. Huang, X. Liu, and F. X. Lin. Approximate query service on autonomous iot cameras. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, page 191–205. Association for Computing Machinery, 2020.

[28] J. Yi, S. Choi, and Y. Lee. Eagleeye: Wearable camera-based person identification in crowded urban spaces. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, New York, NY, USA, 2020. Association for Computing Machinery.