Degradable Inference for Energy Autonomous Vision Applications

Alessandro Montanari[†] **Mohammed Alloulah**[†] **Fahim Kawsar**[†‡]
[†]Nokia Bell Labs, Cambridge, UK [‡]TU Delft, Netherlands
{alessandro.montanari, mohammed.alloulah, fahim.kawsar}@nokia-bell-labs.com

ABSTRACT

Mobile vision systems, often battery-powered, are now incredibly powerful in capturing, analyzing, and understanding real-world events uncovering interminable opportunities for new applications in the areas of life-logging, cognitive augmentation, security, safety, wildlife surveillance, etc. There are two complementary challenges in the design of a mobile vision system today - improving the *recognition accuracy* at the expense of *minimum energy* consumption. In this work, we posit that best-effort sensing with degradable featurization and an elastic inference pipeline offers an interesting avenue to bring energy autonomy to mobile vision systems while ensuring acceptable recognition performance. Borrowing principles from Intermittent Computing, and Numerical Computing we propose such best-effort sensing using a *Degradable-Inference* pipeline supported by a parameterized Discrete Cosine Transformation (DCT) based featurization and an Anytime Deep Neural Network. These two principles aim at extending the lifetime of a mobile vision system while minimizing compute and communication cost without compromising recognition performance. We report the design and early characterization of our proposed solution.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Image compression**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Energy autonomous, anytime algorithms, neural networks.

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3349337

ACM Reference Format:
Alessandro Montanari, Mohammed Alloulah, and Fahim Kawsar. 2019. Degradable Inference for Energy Autonomous Vision Applications. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3341162.3349337

1 INTRODUCTION

The advent of the Internet of Things (IoT)—whereby billions or trillions of devices sense and act in the environment—has catalysed the "Deploy and Forget" vision. With the scaling of IoT devices comes significant effort and cost barriers to mass deployability and maintenance. Energy harvesting sensing systems, also known as battery-less or energy autonomous systems, support and simplify this vision. By taking advantage of freely available energy, mass deployment and maintenance barriers can be potentially eliminated or greatly reduced. Additionally, operating without batteries liberates such systems from safety or environmental concerns and reduces their overall size.

Vision-based tasks are emerging as ambitious use cases of ultra-low power devices. Examples include elderly care [22], spotting the presence of certain wild animals in remote locations [10], cognitive augmentation [14], and intruder detection. While ultra-low power systems are unable to compete on high-fidelity performance metrics, they could act as a scalable cost-effective solution to spotting events of interest and subsequently triggering more expensive scene analyses. Our work herein targets this class of applications. The term *Visual Wake Words* [7] encapsulates such concept and draws parallels with conventional audio hot keyword detection wherein a certain spoken word triggers a fully-fledged speech recognition system.

Clearly, energy autonomous systems are characterized by high uncertainty in power availability; the time at which they are awake and can sense the environment or process incoming data is highly unpredictable. Moreover, given that the energy distribution is not uniform over time, the duration of the awake periods is uncertain too. This behavior is denoted as *intermittent computing* [25]. Enabling vision-based

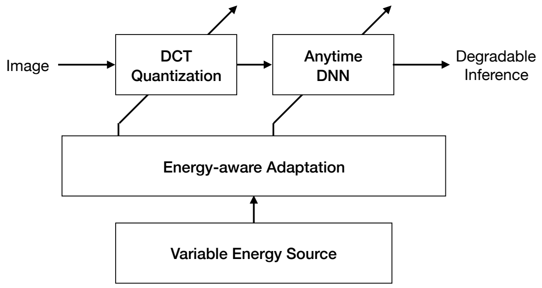


Figure 1: Energy-aware execution whereby inference can be degraded in order to track an intermittent energy envelope.

applications on such constrained devices involves great challenges which we delineate in the next section.

Challenges and Requirements

The two main requirements to enabling practical and useful energy-autonomous vision applications are: (1) low latency response and (2) high system availability. These requirements ensure such systems could be deployed in environments where a prompt response is necessary (e.g., continuous face identification for cognitive assistance). The composite challenge in this context is the optimal use of scarce, sporadic energy while reducing the overhead of the runtime system.

Advancements in the field of *intermittent computing* have mainly focused on the continuation of computation across power failures [5, 8, 29]. Underpinning these important works is the need to allow a computational task to span across multiple power failures, thereby ensuring consistency of the task's internal state. While these works represent essential building blocks to simplifying the development of user applications, significant latency might be introduced between sensing and the resultant action. Another strand of prior work concerns runtime systems that allow developers to define time constraints in order to ensure data timeliness [16]. By means of these time constraints, "stale" data—i.e., data that is no longer relevant to current state of the environment—is discarded as to avoid wasting precious energy and compute.

Contrary to the above state-of-the-art techniques, we consider in this work a class of visual classification applications for which low latency computation is paramount in order to facilitate responsive data consumption within the available energy envelope.

Overview of Proposed Approach

To overcome the limitations of previous work described above, we propose the system architecture depicted in Figure 1. At a high level, the system adapts to a fluctuating energy budget through two control mechanisms: (i) variable Discrete Cosine Transform (DCT) quantization and (ii) an anytime deep neural network (DNN) with many intermediate output stages. Such adaptations are denoted by the diagonal arrows in Figure 1.

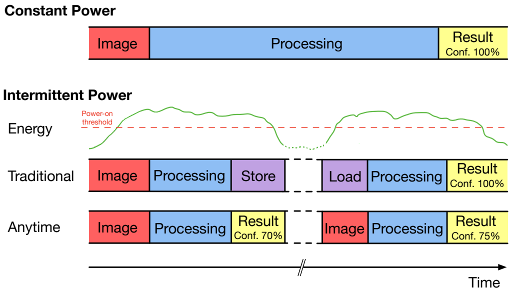


Figure 2: Examples of processing on input images when constant power is available (top) and when energy is harvested from the environment and therefore intermittent (bottom). In traditional intermittent computing systems, a computational task spans across power failures and is complete, delivering a result with 100% confidence. Instead, *degradable inference* produces an approximate result whose confidence is proportional to available energy, thus simultaneously optimizing responsiveness and energy utility.

Given our focus on vision processing, we propose the adoption of Discrete Cosine Transform (DCT) quantization as a tunable pre-processing block that reduces the complexity of subsequent steps. The aim is to marry DCT-based degradable pre-processing to anytime neural network architectures in an end-to-end energy-ware execution framework. The DNN follows the paradigm of anytime algorithms [6] and produces a valid answer even if interrupted before completion. Further, it refines initial results when allocated more time for computations. Such framework would adapt the amount of required computations based on the dynamic energy envelope generated by environmental harvesters.

When a sudden power failure occurs, traditional runtime systems resort to storing the internal state of the task along with intermediate results in non-volatile memory [5, 8, 29]. Upon power resumption, computations are continued and the cycle is repeated until task completion. In contrast, the proposed system matches approximate results to available power budget, thereby guaranteeing a timely response. That is, the system trades off confidence in results for energy and compute efficiency. Figure 2 illustrates this behavior.

In the following, we show how JPEG-style DCT can be used in principle to *gracefully* degrade image representation subject to instantaneous energy budget. We justify our formulation with early feasibility characterization and further posit its merits within the context of highly unpredictable energy autonomous system.

2 ARCHITECTURE

In meeting the challenges of an energy autonomous mobile vision application, we propose a system which relies on two main components to adapt computation in accordance to a variable energy budget: (i) variable DCT quantization and (ii) anytime DNN (Figure 1). We elaborate further on these two subsystems.

(i) DCT quantization. One would think that—in accordance with decades of vision processing domain expertise both in still images (e.g. JPEG) or video (e.g MPEG)—spatial frequencies should come as a natural building block representation for subsequent processing. However, it was not until very recently that the machine learning community has discovered the optimality of DCT as a pre-processor for state-of-the-art image classification DNNs [13]. Gueguen et al. show that a JPEG-style DCT layer applied on images outperforms any other *learnt* alternative both in terms of classification accuracy and speed (i.e., computational efficiency). Such observation motivates us to rely on a JPEG-style DCT stage that feeds a DNN spatial coefficients as shown in Figure 1. This neural pre-processig stage allows us to *sparsify* the input layer of the DNN as a *linear* function of available energy budget. By virtue of DCT linearity, image representation fidelity can be *gracefully* traded off for compute efficiency. We posit that such formulation is of particular benefit to the emerging paradigm of approximate computation.

(ii) Anytime Deep Neural Network. Despite the energy-aware adaptation described above, deep neural networks remain computationally heavy for small micro-controllers. A recent work tackles loop-heavy computations in DNNs as to allow efficient forward execution under power failures [12]. However, when computations are interrupted, no valid output is produced by the network because data can no longer traverses the remainder of the computational graph. Consequently, additional latency in incurred until another charge cycle would allow the resumption and completion of the computational task. To overcome this fundamental limitation, we propose to utilize a DNN which conforms to the anytime paradigm [6] through the dynamic rerouting of the computational graph subject to available energy budget.

Specifically, a class of DNNs in machine learning is designed to output lower-fidelity inferences at certain intermediate layers in addition to the final deepest output inference [18, 28]. As a byproduct, these lower-fidelity intermediate inferences represent a controlled grading of performance as a function of feature embedding complexity/sophistication. These models, despite the enhancing inference efficiency, remain hard to port to devices with stringent resource constraints. Careful co-design of inference models and the underlying hardware is necessary to enable efficient execution on ultra-low power devices. We leave this to future work.

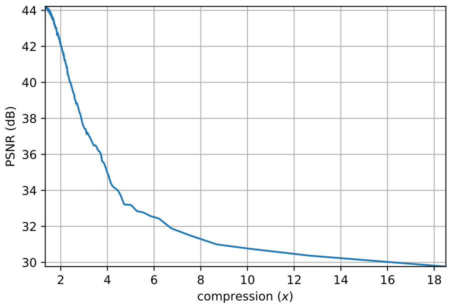


Figure 3: Example compression versus image distortion curve arrived at by varying quantization levels in JPEG's 2D cosine transform.

We dub the combined linear and nonlinear energy-ware operation *degradable inference*. Next section provides early investigations designed to shed more light on variable DCT quantization as a viable mechanism for energy-aware adaptation.

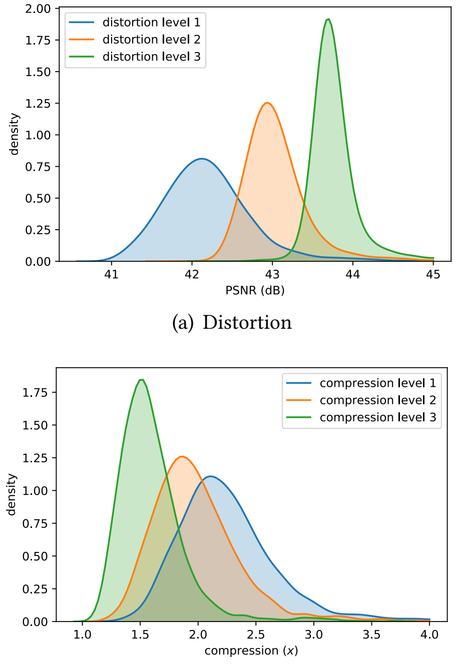
3 PRELIMINARY RESULTS

In this section we report initial results regarding the two components of our proposed system: input DCT quantization and anytime deep neural network, as well as their combination in a preliminary degradable inference pipeline.

DCT Quantization

Method. We use CIFAR10 image dataset [23]. We implement a standard JPEG image compression pipeline in Python. In order to characterize the amount of distortion introduced by variable DCT quantization, we implement the following. First, an RGB image is converted to its YCbCr representation. Second, variable DCT quantization is applied as per the JPEG standard and the amount of compression is calculated i.e. ratio of remaining coefficients to full image fidelity. Third, the quantized DCT coefficients are converted to YCbCr and then back to RGB. Forth, the Peak Signal-to-Noise Ratio [27] (PSNR) between the round-trip image reconstruction and the original image is computed. Finally, a compression-distortion curve is arrived at as shown in Figure 3 by repeating for a range of quality factors during DCT quantization. In order to study the statistics of this procedure, a subset of 5000 images from CIFAR10 dataset is processed as outlined.

Results. The distributions of image quality (PSNR) for three equally-spaced quantization levels are shown in Figure 4(a) across the CIFAR10 subset. Specifically, these quantization levels gradually and systematically seem to *gracefully* control PSNR. The corresponding compression distributions at the same quantization levels are shown in Figure 4(b). Remember compression here equates to input layer sparsification for subsequent DNN. That is, 2× compression sparsifies input



(a) Distortion

(b) Compression

Figure 4: Examples of 3 DCT quantization levels gracefully controlling amount of compression and distortion.

activations by half, which proves that the DCT linear "nob" is in principle capable of *gracefully* degrading input subject to instantaneous energy budget.

In order to further study the fine-grained distortion- compression trade-off, we turn to the joint 2D density of Figure 5. Specifically across the CIFAR10 subset, greater than 40% sparsification of input is achieved while maintaining a PSNR in excess of 43 dB. Further sparsification gains are possible with a 41+ PSNR. These early findings suggest that it is possible to build sparsification profiles at design-time that are able track dynamic energy budgets whilst gracefully degrading input quality. Further validation on the inference performance is ongoing.

Anytime DNN

Method. We build a ResNetv2 model [15] with 11 layers (10 convolutional and 1 dense, together with batch normalization and average pooling). Inspired by the BranchyNet architecture [28], we add two additional branches at progressively deeper positions in the network. The first branch adds two convolutional layers and one dense layer, while the second branch has one convolutional and one dense layer. Thus we obtain a model with 3 outputs (or stages). The network, therefore, is "anytime" because even when a prediction is computed at the first output branch, computations could continue—progressively building on more complex features—to produce a refined prediction at the second output branch, and/or at the third branch.

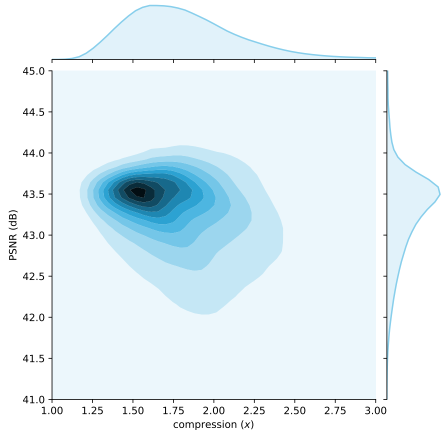


Figure 5: Joint compression-distortion density across a subset of CIFAR10 images.

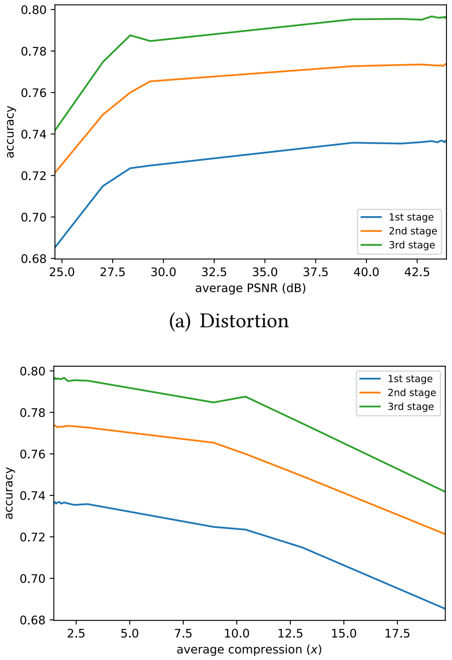
Table 1: Accuracy on the CIFAR10 test set and estimated number of floating point operations (FLOPS) for each output of the model.

Output	Accuracy	# FLOPS
Stage 1	73%	263k
Stage 2	77%	336k
Stage 3	80%	668k

Results. We train and test the above model on the CIFAR10 dataset, using the Adam optimizer [21] and data augmentation during training. Table 1 shows how the accuracy of the model progressively increase at deeper outputs. Table 1 also lists the computational workloads in flops associated with deeper network stages. We leave for future work the design of a more sophisticated model, evaluation on bigger and more realistic datasets (e.g. ImageNet) and the implementation on an energy autonomous system.

Degradable Inference

Method. In an experiment aimed at early concept validation, we combine a variable JPEG-style DCT quantization with a 3-stage Anytime DNN. This particular DNN operates on raw RGB pixels. However, we emulate the effect of DCT input layer sparsification through a round-trip quantization in the spatial frequency domain. Specifically, for a CIFAR10 testset, we sweep the quantization quality factor as per the JPEG standard [19, 26] in order to effect a controlled image degradation. We proceed to feed the Anytime DNN with



(a) Distortion

(b) Compression

Figure 6: Accuracy of a 3-stage Anytime DNN under variable DCT-based input sparsification.

the degraded CIFAR10 testset corresponding to these quantization levels on a test-by-test basis. We then measure the classification performance at the three DNN stages.

Results. Figure 6 shows the accuracy of the three DNN stages as a function of DCT quantization distortion in 6(a) and compression in 6(b). For instance, it is evident from Figure 6(b) that as we proceed to more aggressively sparsify the input image DCT representation, we obtain proportional inference degradation across the three DNN output stages, consistently. The trend is equivalently conveyed by the distortion-accuracy trade-offs of Figure 6(a). These preliminary results, despite having been generated using a round-trip emulated controlled grading, serve to highlight the feasibility of our proposal, subject to adapting the DNN network to operate directly on the DCT image representation as opposed to raw RGB pixels. Recent prior art not only confirms the feasibility of this DCT approach, but also uncovers its optimality vis-à-vis DNNs trained on raw RGB pixels [13].

4 OUTLOOK

Low power micro-controllers are the preferred choice when low cost computation is required, for example in industrial or smart-home applications. Further, it has recently become feasible to use low power micro-controllers for traditional computer vision tasks [3] or to run deep learning models [1, 2]. That said, the combined effect of limited compute and memory resources on such small chips presents great challenges for computer vision tasks. As touched on earlier, the intermittent behaviour of energy harvester systems further compounds these challenges [25]. In such peculiar environment, ensuring the forward progression of computational workloads despite power failures, as tackled by several previous works [5, 8, 29], is crucial.

In this paper, we propose a complementary approach which adapts computations to the dynamic energy budget (or envelope) following the *anytime* paradigm [6]. The central premise behind anytime algorithms is the ability to be interrupted before workload completion—say as a result of energy depletion—while retaining the ability to output a valid, albeit degraded, result. Further, a wider energy envelope allows anytime algorithms to progressively refine their earlier approximate result. Anytime algorithms have been originally used for time-dependent planning and decision-making [9, 17, 24]. They are particularly suited for applications with stringent real-time requirements which can tolerate lower accuracy but where response time is paramount. To the best of our knowledge, the only work to apply this concept to energy harvesting systems is [11]. In this work, Ganesan et al. propose a hardware-software co-design approach to process data at a subword granularity. This enables approximate computation of a certain family of operations which can be further improved if the entire data words are used for the computation. However, this work requires hardware modifications of existing processors and its applicability is limited to specific operations. Outside intermittent computing, a similar notion of *degradable* coding has been proposed in wireless communications whereby video fidelity *gracefully* tracks instantaneous channel conditions [4, 20].

In contrast to prior art, we focus on a vision application reliant on two software-only subsystems, which together provide control mechanisms to adjust the amount of computation in order to track a dynamic energy envelope. The idea is, in part, motivated by recent advancements in machine learning research that demonstrate the optimality of established vision coding techniques from expert domains (e.g. JPEG and MPEG) over any learnt approach operating on raw pixels [13]. Through a preliminary investigation we characterize the amount of distortion introduced by variable Discrete Cosine Transform (DCT) quantization and show how in principle it is capable of *gracefully* degrading input subject to instantaneous energy budget. Additionally, we demonstrate how variable input quantization, emulated through a round-trip DCT compression, can be combined with an anytime deep neural network, as to result in a two-“nob” control for inference degradability.

Future work will adapt the DNN model to work directly on the DCT coefficients, instead of the raw RGB pixels. We will also investigate interaction between training and levels of input degradation. We will further improve the anytime network in order to achieve better performance and generalizability to larger datasets. The ultimate objective is to demonstrate an end-to-end system whereby an embedded device utilizes a combination of energy harvesters (e.g., solar and radio) and ultra-low power communication techniques in order to communicate degradable inferences to a remote intermittent host system.

REFERENCES

[1] 2019. Basic MNIST handwriting recognition with uTensor. https://github.com/uTensor/utensor-mnist-demo. Accessed: 2019-06-29.

[2] 2019. Image recognition on Arm Cortex-M with CMSIS-NN. https://developer.arm.com/solutions/machine-learning-on-arm/developer-material/how-to-guides/image-recognition-on-arm-cortex-m-with-cmsis-nn. Accessed: 2019-06-29.

[3] 2019. OpenMV Cam H7. https://openmv.io/collections/products/products/openmv-cam-h7. Accessed: 2019-06-29.

[4] Siripuram Aditya and Sachin Katti. 2011. FlexCast: Graceful wireless video streaming. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 277–288.

[5] Domenico Balsamo, Alex S Weddell, Geoff V Merrett, Bashir M Al-Hashimi, Davide Brunelli, and Luca Benini. 2014. Hibernus: Sustaining computation during intermittent supply for energy-harvesting systems. *IEEE Embedded Systems Letters* 7, 1 (2014), 15–18.

[6] Mark Boddy and Thomas L Dean. 1989. *Solving time-dependent planning problems*. Brown University, Department of Computer Science.

[7] Aakanksha Chowdhery, Pete Warden, Jonathon Shlens, Andrew Howard, and Rocky Rhodes. 2019. Visual Wake Words Dataset. *arXiv preprint arXiv:1906.05721* (2019).

[8] Alexei Colin and Brandon Lucia. 2016. Chain: tasks and channels for reliable intermittent programs. In *ACM SIGPLAN Notices*, Vol. 51. ACM, 514–530.

[9] Thomas L Dean and Mark S Boddy. 1988. An Analysis of Time-Dependent Planning.. In *AAAI*, Vol. 88. 49–54.

[10] Andy Rosales Elias, Nevena Golubovic, Chandra Krintz, and Rich Wolski. 2017. Where's the bear? Automating wildlife image processing using iot and edge cloud systems. In *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 247–258.

[11] Karthik Ganesan, Joshua San Miguel, and Natalie Enright Jerger. 2019. The What's Next Intermittent Computing Architecture. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 211–223.

[12] Graham Gobieski, Brandon Lucia, and Nathan Beckmann. 2019. Intelligence beyond the edge: Inference on intermittent embedded systems. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 199–213.

[13] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. 2018. Faster neural networks straight from JPEG. In *Advances in Neural Information Processing Systems*. 3933–3944.

[14] Kiryong Ha, Zhuo Chen, Wenlu Hu, Wolfgang Richter, Padmanabhan Pillai, and Mahadev Satyanarayanan. 2014. Towards wearable cognitive assistance. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 68–81.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.

[16] Josiah Hester, Kevin Storer, and Jacob Sorber. 2017. Timely execution on intermittently powered batteryless sensors. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 17.

[17] Eric J Horvitz. 2013. Reasoning about beliefs and actions under computational resource constraints. *arXiv preprint arXiv:1304.2759* (2013).

[18] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. 2017. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844* (2017).

[19] ISO IEC. 1994. Information technology-digital compression and coding of continuous-tone still images: Requirements and guidelines. *Standard, ISO IEC* (1994), 10918–1.

[20] Szymon Jakubczak and Dina Katabi. 2011. A cross-layer design for scalable mobile video. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 289–300.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Krishna Reddy Konda, Andrea Rosani, Nicola Conci, and Francesco GB De Natale. 2014. Smart camera reconfiguration in assisted home environments for elderly care. In *European Conference on Computer Vision*. Springer, 45–58.

[23] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.

[24] Victor R Lesser, Jasmina Pavlin, and Edmund Durfee. 1988. Approximate processing in real-time problem solving. *AI magazine* 9, 1 (1988), 49–49.

[25] Brandon Lucia, Vignesh Balaji, Alexei Colin, Kiwan Maeng, and Emily Ruppel. 2017. Intermittent computing: Challenges and opportunities. In *2nd Summit on Advances in Programming Languages (SNAPL 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[26] Michael Robin and Michel Poulin. 1997. *Digital television fundamentals*. McGraw-Hill New York.

[27] David Salomon. 2013. *A guide to data compression methods*. Springer Science & Business Media.

[28] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2464–2469.

[29] Kasım Sinan Yıldırım, Amjad Yousef Majid, Dimitris Patoukas, Koen Schaper, Przemyslaw Pawelczak, and Josiah Hester. 2018. Ink: Reactive kernel for tiny batteryless sensors. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 41–53.