# Onboard Multi-Scale Tile Classification for Satellites and Other Spacecraft

Evan W. Gretok, Alan D. George
*NSF Center for Space, High-performance, and Resilient Computing (SHREC)*
*Department of Electrical and Computer Engineering*
*Swanson School of Engineering, University of Pittsburgh*
Pittsburgh, Pennsylvania, United States of America
{evan.gretok, alan.george}@nsf-shrec.org

2021 IEEE Space Computing Conference (SCC) | 978-1-6654-2400-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/SCC49971.2021.00019

***Abstract*—As space-based sensors for Earth observation continue to increase in resolution, extreme computational capability is required to process, compress, and interpret the exorbitant volumes of data collected. Historically, computation has been performed on ground systems. However, this research proposes and benchmarks a flexible onboard tile-classification system for high-resolution Earth observation on satellites and spacecraft. Modern computer-vision techniques for classification and segmentation have progressed to orbital platforms. However, for many applications, the granularity of data analysis has tradeoffs: classification of the entire image is too coarse to yield useful scientific products, and segmentation at the pixel-level is too computationally expensive. Tile classification serves as a middle ground between these two paradigms that can be applied to high-resolution imagery. Transfer learning is conducted on large and small variants of MobileNetV2 deep-learning models using Earth-observation imagery at different ground-resolved distances (GRDs). A case study then compares the inference accuracy for fine-tuned models tested on different GRDs. This process demonstrates the effectiveness of applying one model to a variety of image scales. The inference performance for these models is evaluated in terms of execution time, parallel performance, memory use, and energy consumption. This research showcases increased capacity for onboard remote sensing with tile classification that can foster more versatile space situational awareness.**

***Index Terms*—Space Computing, Onboard Processing, Image Tile Classification, Remote Sensing, Autonomy, Deep Learning, Image Processing, Computer Vision**

## I. Introduction

Navigation, communication, science, and surveillance have all been facilitated from low-Earth orbit for decades. More recently, vast increases in compute capability have made space a new proving ground for intelligence, awareness, and autonomy. Constellations of distributed small satellites (SmallSats) image the Earth's entire surface each day, thus there is tremendous overhead in generating actionable data. Advances in optics, sensors, and processors now enable ground-resolved distances (GRDs), physical distances between digital pixels in an image, at the decimeter scale. Single images captured by satellites can contain thousands of individual people, vehicles, buildings, and other features of interest. Additionally, separate sections of an image may need to be processed in different ways to extract relevant data. Modern computer-vision techniques can perform classification, object detection, segmentation, and other tasks to facilitate this need. This research seeks to perform processing onboard, which can allow for decentralized, intelligent aggregation of data and enable satellites and spacecraft to actively formulate decisions and prioritize actions.

With sensor resolution increasing, optical and algorithmic advances reducing GRD, demand for data expanding, and the number of Earth-observation satellites continuing to grow, manual human-in-the-loop operation is becoming impractical. Considering the rapidly changing conditions captured by space-based remote sensing, any delay in translating raw information to actionable data may thwart an effective response. The transition from expensive, monolithic satellites to agile constellations of SmallSats also includes challenges and tradeoffs. While capacity for imaging grows and satellites of a constellation may observe many locations simultaneously, the overhead required to manage and maintain the constellation and the data it outputs can be daunting. There is opportunity, though, in the independent function of dozens of Earth-observation systems. Each satellite could possess sufficient processing capability to evaluate the data it ingests onboard, analyze and interpret actionable information, and autonomously formulate a mission operation in response. However, common Earth-observation satellites may vary widely in GRD, from 0.3 m for WorldView, to 3.0 m for Planet's Dove, to 30 m for Landsat 8 and 9. Many datasets provide ample training material for low-GRD land-cover classification but are not as useful at higher GRDs for which smaller features are not visible. This study arrives at a selection of classes with features that scale effectively to multiple GRDs.

The research proposes an onboard, high-resolution tile-classification system for Earth observation that is benchmarked for runtime, parallel performance, memory use, and energy consumption. The accuracy of training and applying models at different GRDs is explored. The effect of transfer learning on small datasets, especially considering simple means of manual augmentation for rotationally invariant satellite imagery, is considered. This tile-classification system is intended to serve as a demonstration of advanced space-based capabilities in

This research was supported by the NSF Center for Space, High-performance, and Resilient Computing (SHREC) industry and agency members and by the IUCRC Program of the National Science Foundation under Grant No. CNS-1738783.

978-1-6654-2400-4/21/$31.00 ©2021 IEEE
DOI 10.1109/SCC49971.2021.00019

computer vision for onboard sensor-data processing.

This paper is structured as follows. Section II details important background on space computing, remote sensing, the models employed, and the imaging and computing platforms considered. Section III discusses the methods employed in this research, including dataset creation, model transfer learning, inter-GRD comparison of models, and the benchmarking process. Section IV presents results from each section of the study. Section V offers discussion and analysis of trends. Section VI defines future directions for this research. Section VII closes with a summary and key highlights.

## II. Background and Related Research

This section provides background information on relevant concepts and covers numerous related studies. An overview of recent efforts in space computing, especially deep learning in space, is provided. The models and platforms employed in this study are also discussed.

### A. Space Computing and Autonomy

Space-computing capabilities vary based on the platform employed. Radiation-hardened (rad-hard) space processors are virtually immune to many hazards of space radiation with minimal risk of damage or data corruption [1]–[3]. Radiation-tolerant (rad-tol) space processors typically combine commercial-off-the-shelf (COTS) and rad-hard components to enable high-performance, energy-efficient processing capability with dependable monitoring and management subsystems to ensure reliable operation [4], [5]. COTS processors offer substantially greater performance compared to rad-hard alternatives but are susceptible to radiation effects that can impact system reliability. As compute capability has increased, onboard machine-learning and artificial-intelligence algorithms have become amenable to both COTS and next-generation rad-hard platforms alike. [6] features an NVIDIA Jetson TX1 performing space image-processing and computer-vision tasks and describes a process for model training and deployment. A simpler approach with a more specific goal, [7] uses a multi-layer perceptron for wildfire detection and attains 99% accuracy. Deep-learning models are demonstrated at up to 11 frames-per-second (FPS) on a rad-tol ARM Cortex-A9 CPU platform in [8]. [9] details an approach for onboard ship detection in low-GRD imagery using the NVIDIA Jetson TX2. Many of these studies include benchmarks demonstrating onboard processing viability.

Rovers, with numerous science objectives and delayed communication due to distance from Earth, are a prime target for onboard autonomy and deep learning. The Mars Exploration Rovers used the Autonomous Exploration for Gathering Increased Science (AEGIS) system for imaging key targets based on provided science objectives. Despite its effectiveness, it was a more classical computer-vision pipeline consisting of several common algorithms [10]. [11] demonstrates an intelligent, ground-based system to identify terrain hazards and recommend safe navigation paths. Success with self-supervised learning on rovers is demonstrated in [12], which focused on terrain-type classification. Additional novelty is found in its co-training of two classifiers on each other's output. This research attained 82% accuracy with only three labelled images on a test rover in a Mars-like environment. [13] identifies this project as the Soil Property and Object Classification (SPOC) system, a deep convolutional neural network (DCNN) capable of classifying 17 different terrain types. [14] proposes the automated Science Captioning of Terrain Images (SCOTI) system to address rover bandwidth limitations and intelligent image prioritization. Intelligent processing systems for rovers and probes are typically based on rad-hard processors, limiting their computational capability. The classification system devised in this research is designed to address remote-sensing and spacecraft-autonomy needs for both satellite and rover applications.

### B. Remote Sensing

Deep-learning-based computer-vision paradigms have been extensively applied to remote-sensing applications for Earth-observation imagery from satellites. [15] details a parallel-branched DCNN for vehicle identification from orbit. [16] shows that DCNNs trained on general image datasets transfer well to aerial photography but, at the time of the study, may be outperformed by color descriptors for space-based remote-sensing tasks. However, it is important to note that low-GRD Earth-observation imagery may be more similar in scale and content to the aerial photography used in that research. [17] presents a land-use classification system employing CaffeNet and GoogLeNet to attain 91.8% accuracy, though the system is deployed on a mobile GPU, which may see lower reliability in a space environment. A tile-based approach focusing on segmentation is shown in [18], and it provides much insight into the complex problem of classifying artifacts on tile edges. [19] demonstrates a variation on a common visual bag-of-words approach by handling local and global features differently. For high-resolution, low-GRD images, [20] provides a thorough survey of semi-supervised classification accuracy on a variety of deep-learning models. Similarly, [21] provides a survey of remote-sensing datasets and classification methods. This study also proposes its own large and effective NWPU-RESISC45 dataset, which attains up to 90.36% on a fine-tuned VGG-16 DCNN. [22] develops and demonstrates BiMobileNet, incorporating architectural changes for improved classification accuracy on remote-sensing imagery. Onboard deployment or performance is not often a focus of these studies, and inference is typically conducted on ground systems using downlinked data. The research described herein aims to make such efforts amenable to onboard processing and provide a path to actionable data and autonomy.

### C. Earth-Observation Platforms

Numerous high-resolution, space-based Earth-observation platforms have been providing imagery for decades. Landsat 8, launched in 2013, provides GRDs of 30 meters per pixel for visible spectra, 15 meters per pixel for panchromatic, and 100 meters per pixel for thermal. This satellite alone contributes

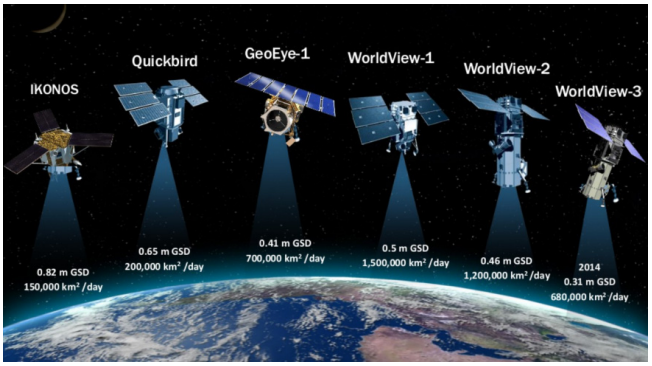


Fig. 1. Earth Observation Satellites [23]

over 500 scenes per day to the United States Geological Survey archives [24]. Landsat 9 will feature the same GRDs for compatibility with the existing dataset and is targeting a 2021 launch [25]. The QuickBird-2 satellite, launched in 2001, provided 2.4-meter multispectral and 0.6-meter panchromatic imagery [26]. WorldView-1, launched in 2007, demonstrated a 0.5-meter panchromatic GRD with a 1.7-day revisit time [27]. WorldView-2 added 1.8-meter multispectral imaging. WorldView-3 expanded infrared capability and further reduced panchromatic GRD to 0.31 meter [28]. WorldView-4, also known as GeoEye-2, provides a similar 0.31-meter panchromatic GRD as well as 1.24-meter red-green-blue (RGB) and near-infrared (NIR) imaging with revisit times as low as five hours. Featuring star trackers, a global positioning system, control moment gyroscopes, and 3.2 Tb of solid-state storage, it is capable of imaging without directives from the ground to within four-meter geolocation accuracy [29]. A collective depiction of many of these platforms is shown in Fig. 1. It is relevant to note that many of these systems do not possess the same standards of accuracy and reliability as Landsat. There are often tradeoffs to such low GRDs, but this is representative of the shifting paradigms of Earth-observation satellites and techniques. A simplified visualization depicting pixel sizes at different GRDs is included in Fig. 2. Houses and vehicles are included to demonstrate detail loss at higher GRDs.

The University of Pittsburgh NSF Center for Space, High-performance, and Resilient Computing (SHREC) operates several ISS-based Earth-observation and space-computing platforms. One of these platforms, the Space Test Program - Houston 6 - Space Supercomputing for Image and Video Processing (STP-H6-SSIVP) experiment, has captured a dataset of several thousand five-megapixel Earth-observation images [30]. SSIVP's GRD is limited to approximately 60 meters per pixel, an even higher GRD than Landsat imagery used in this study and suitable to fewer applications. The STP-H7-CASPR (Configurable and Autonomous Sensor Processing Research) experiment will provide imagery at a roughly 2.0-meter GRD with novel super-resolution algorithms [31]. This platform serves as a target for deployment of this research. These platforms are intended to demonstrate compute capability within SmallSat constraints, but many of the techniques applied can be transferred to larger systems.

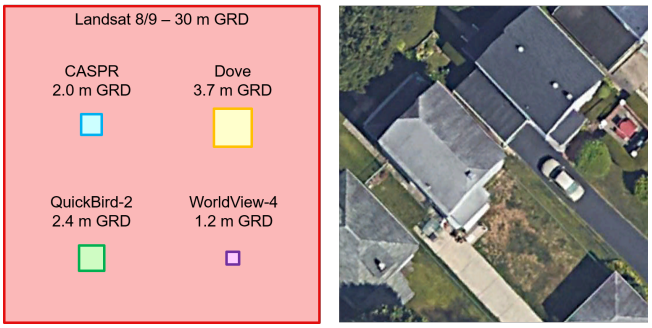


Fig. 2. Visualization of Pixel Sizes at Different GRDs

Planet, a commercial entity specializing in frequent-revisit satellite imagery and data analytics, offers unique capabilities through their constellations [32]. Over 130 Dove SmallSats produce 200 million square kilometers per day of four-band, RGB and NIR, imagery at 3.7 meters per pixel ground sample distance (GSD). The images are orthorectified to correct for distortions from terrain [33]. 21 SkySat satellites image 400,000 square kilometers per day and produce five-band RGB/NIR and panchromatic imagery. These platforms can attain 0.5-meter GSD and offer sub-daily off-nadir and sub-weekly nadir revisit times [34]. Planet maintains a "living database" of its imagery for data analytics, such as the deep-learning model training conducted in this study. Extensive research has been conducted based on this imagery. [35] uses a hybrid CNN and random forest approach for determination of airborne particulate concentration using low-GRD satellite imagery alone. [36] uses Planet imagery for detection and enforcement of illegal fishing. Planet has partnered with SpaceNet for the SpaceNet 7 Multi-Temporal Urban Development Challenge [37] to track changes over time for urban and infrastructure development as well as aid in disaster response [38]. Planet is one example of a modern disruptor in the Earth-observation and remote-sensing paradigms. Commercial, distributed platforms now offer high-resolution and low-GRD imaging along with the underlying infrastructure to make effective use of actionable data. Ultimately, this research is designed to be versatile and applicable to a variety of observation platforms with varying GRDs.

### *D. Deep-Learning Models*

Since the emergence of DCNNs as a solution to image-classification problems with AlexNet in 2012 [39], numerous networks and techniques have claimed state-of-the-art status. This section will consider just a few, mainly those small enough to deploy at reasonable inference speeds on space systems. ResNet increased network depth to improve accuracy, but maintains several smaller network variants, such as ResNet-50 and ResNet-18, which would be amenable to space [40]. The SqueezeNet model demonstrates comparable accuracies to AlexNet with 50 times fewer parameters and can be compressed to less than a megabyte in size [41].

TABLE I
PLATFORM SPECIFICATIONS

| Platform | System-on-Chip | Architecture | Tested Frequency (MHz) | Execution Units | Peak Power (W) | Memory (GB) | Memory Type | Memory Frequency (MHz) |
|---|---|---|---|---|---|---|---|---|
| PYNQ-Z2 (SSP) | Zynq 7020 | ARM Cortex-A9 | 650 | 2 | 5 | 0.5 | DDR3 | 525 |
| HiKey LeMaker (HPSC Full) | Kirin 620 | ARM Cortex-A53 | 729 | 8 | 15 | 2.0 | LPDDR3 | 800 |
| ODROID-C2 (HPSC Half) | AmLogic S905 | ARM Cortex-A53 | 1000 | 4 | 7 | 2.0 | DDR3 | 912 |
| Raspberry Pi 4 (8QuadMax) | Broadcom BCM2711 | ARM Cortex-A72 | 1500 | 4 | 5 | 4.0 | LPDDR4 | 1600 |

DenseNet employs interconnection between layers, reducing the effect of the vanishing gradient problem and achieving high accuracy with efficient computation and memory use [42]. MobileNetV2, aptly named, is optimized for mobile devices using advanced architectural techniques, inverted residuals and linear bottlenecks, to reduce latency and parameter count [43]. MobileNetV3 improves efficiency with adjustment of underlying operations and further optimizes the model for hardware accelerators, such as Google's Edge TPU [44]. GhostNet further improves on MobileNetV3 performance and accuracy by extracting additional features using computationally inexpensive operations [45]. [46] adjusts the GhostNet architecture to reduce overhead but maintains high accuracy in tests with remote-sensing data. While the model employed is a key determinant of performance, this research can be extended to a variety of novel deep-learning classification architectures.

### *E. Datasets*

Several key datasets have been adopted as standard for use with remote-sensing applications and Earth-observation imagery. The Brazilian coffee dataset, presented in [16], is a simple binary classification dataset to determine the presence or absence of a coffee crop in remote-sensing imagery. The UC Merced dataset, presented in [47], provides 2100 images from the United States Geological Survey National Map sorted into 21 classes. Some UC Merced classes, such as airplanes, sports fields, marinas, and overpasses, may represent features that are too small to be relevant at the high GRDs that are considered in this study. Other classes, like agriculture, forest, residential and commercial areas, rivers, and coastlines, should still be amenable to a variety of GRDs though scaled to differing sizes. The NWPU-RESISC45 dataset presented in [21] offers many more images and classes, 31,500 photos sorted into 45 categories. While the NWPU-RESISC45 images range in GRD from 0.2 meters to 30 meters, the vast majority of the images represent single-digit GRDs. Many of the new classes represent a familiar dilemma: they are features too small or infrequent to be transferable to a variety of GRDs. This study will seek to construct a set of relevant classes that can be more effective at all GRDs. However, it will also focus on datasets optimized to a single GRD, keeping classes and feature sizes fixed as well as reducing confusion induced by scale variance.

### *F. Computing Platforms*

This study focuses on four specific space-computing platforms based on ARM Cortex-A architectures. The CHREC Space Processor (CSP) [4], [5] and SHREC Space Processor (SSP) are both rad-tol platforms that employ a hybrid combination of COTS components for performance and energy-efficiency as well as rad-hard components for reliability. Both CSP and SSP are based on the dual-core ARM Cortex-A9 architecture in the form of Xilinx Zynq-7000 system-on-chips (SoCs). This architecture will be represented as a TUL PYNQ-Z2 development board with the processing system clocked to 650 MHz [48]. Numerous missions have demonstrated the flightworthiness of the Zynq platform [5], [30], [49], [50]. The High-Performance Spaceflight Computing (HPSC) processor is the culmination of effort by NASA and the United States Air Force Research Lab (AFRL) to meet the increasing need for space-based computation. While currently still in development, the current design of the HPSC will feature a rad-hard high-performance cluster with dual interconnected quad-core ARM Cortex-A53 processors at 800 MHz [51]. The HiKey LeMaker development board employs a similar SoC architecture with two interconnected quad-core ARM Cortex-A53 processors and is thus featured as an emulation platform for the HPSC [52]. The HiKey development board will be clocked to 729 MHz in this research, the nearest supported frequency to the HPSC's projected maximum. The Hardkernel ODROID-C2 development board also employs an ARM Cortex-A53 architecture processor, but consists of only four cores, half that of a full HPSC chiplet [53]. Despite this, it will be used to support additional data collection for comparison and will be clocked to 1000 MHz, its nearest supported frequency to the HPSC's 800 MHz. The ARM Cortex-A72 architecture is being considered in recent mission developments, with the NXP i.MX 8QuadMax silicon-on-insulator rad-tol variant bringing enhanced reliability at speeds up to 1.6 GHz [54]. This platform is represented by the Raspberry Pi 4 clocked at the default 1.5 GHz, the nearest supported frequency to the 8QuadMax platform [55]. A table of platform specifications is visible in Table I.

## III. APPROACH

This section details the methods used in conducting this research. The procedures for dataset assembly, model transfer learning, and the inter-GRD comparison study are detailed. Steps for model inference benchmarking on each platform are noted.

### *A. Dataset Assembly*

The UC Merced and NWPU-RESISC45 imagery datasets were acquired from their respective sources [21], [47]. A manual training and testing split was created for each of these two datasets. 80% of the images were relegated to training and

TABLE II
CLASS REFINEMENT

| NWPU-RESISC45 Class | Feature Target | Max Feature Size (m) | Pixel Width (px) for GRD | | | Percent of Tile (%) | | | Status |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.3 m | 3.0 m | 30.0 m | 0.3 m | 3.0 m | 30.0 m | |
| Airplane | Widest Point | 71.74 | 239.13 | 23.91 | 2.39 | 106.76% | 10.68% | 1.07% | Excluded |
| Airport | Diagonal Length | 5400.07 | 18000.23 | 1800.02 | 180.00 | 8035.82% | 803.58% | 80.36% | Included |
| Baseball Diamond | Diagonal Length | 52.18 | 173.93 | 17.39 | 1.74 | 77.65% | 7.76% | 0.78% | Excluded |
| Basketball Court | Diagonal Length | 50.66 | 168.87 | 16.89 | 1.69 | 75.39% | 7.54% | 0.75% | Excluded |
| Beach | Widest Point | 192.22 | 640.73 | 64.07 | 6.41 | 286.04% | 28.60% | 2.86% | Included* |
| Bridge | Widest Point | 157.71 | 525.70 | 52.57 | 5.26 | 234.69% | 23.47% | 2.35% | Excluded |
| Chaparral | No Dataset Presence | - | - | - | - | - | - | - | Excluded |
| Church | Diagonal Length | 202.63 | 675.43 | 67.54 | 6.75 | 301.53% | 30.15% | 3.02% | Excluded |
| Circular Farmland | Diameter | 1034.82 | 3449.40 | 344.94 | 34.49 | 1539.91% | 153.99% | 15.40% | Merged |
| Cloud | Diagonal Length (Scale) | - | - | - | - | - | - | - | Included |
| Commercial Area | Diagonal Length | 3119.91 | 10399.70 | 1039.97 | 104.00 | 4642.72% | 464.27% | 46.43% | Included |
| Dense Residential | Diagonal Length | 4966.48 | 16554.93 | 1655.49 | 165.55 | 7390.60% | 739.06% | 73.91% | Included |
| Desert | Diagonal Length (Scale) | - | - | - | - | - | - | - | Included |
| Forest | Diagonal Length (Scale) | - | - | - | - | - | - | - | Included |
| Freeway | Widest Point | 87.91 | 293.03 | 29.30 | 2.93 | 130.82% | 13.08% | 1.31% | Excluded |
| Golf Course | Max Hole Length | 496.67 | 1655.57 | 165.56 | 16.56 | 739.09% | 73.91% | 7.39% | Excluded |
| Ground Track Field | Diagonal Length | 203.97 | 679.90 | 67.99 | 6.80 | 303.53% | 30.35% | 3.04% | Excluded |
| Harbor | Diagonal Length | 738.34 | 2461.13 | 246.11 | 24.61 | 1098.72% | 109.87% | 10.99% | Excluded |
| Industrial Area | Diagonal Length | 1016.00 | 3386.67 | 338.67 | 33.87 | 1511.90% | 151.19% | 15.12% | Merged |
| Intersection | Diagonal Length | 56.07 | 186.90 | 18.69 | 1.87 | 83.44% | 8.34% | 0.83% | Excluded |
| Island | Diagonal Length (Scale) | 4308.01 | 14360.03 | 1436.00 | 143.60 | 6410.73% | 641.07% | 64.11% | Included |
| Lake | Diagonal Length (Scale) | 35291.02 | 117636.73 | 11763.67 | 1176.37 | 52516.40% | 5251.64% | 525.16% | Included |
| Meadow | Diagonal Length (Scale) | - | - | - | - | - | - | - | Included |
| Medium Residential | Diagonal Length | 2724.51 | 9081.70 | 908.17 | 90.82 | 4054.33% | 405.43% | 40.54% | Included |
| Mobile Home Park | Diagonal Length | 710.98 | 2369.93 | 236.99 | 23.70 | 1058.01% | 105.80% | 10.58% | Excluded |
| Mountain | Diameter | 4180.82 | 13936.07 | 1393.61 | 139.36 | 6221.46% | 622.15% | 62.21% | Included |
| Overpass | Widest Point | 90.27 | 300.90 | 30.09 | 3.01 | 134.33% | 13.43% | 1.34% | Excluded |
| Palace | Diagonal Length | 206.81 | 689.37 | 68.94 | 6.89 | 307.75% | 30.78% | 3.08% | Excluded |
| Parking Lot | Diagonal Length | 980.06 | 3266.87 | 326.69 | 32.67 | 1458.42% | 145.84% | 14.58% | Excluded |
| Railway | Widest Point | 112.18 | 373.93 | 37.39 | 3.74 | 166.93% | 16.69% | 1.67% | Excluded |
| Railway Station | Diagonal Length | 146.00 | 486.67 | 48.67 | 4.87 | 217.26% | 21.73% | 2.17% | Excluded |
| Rectangular Farmland | Diagonal Length | 9565.20 | 31884.00 | 3188.40 | 318.84 | 14233.93% | 1423.39% | 142.34% | Included |
| River | Widest Point | 5799.98 | 19333.27 | 1933.33 | 193.33 | 8630.92% | 863.09% | 86.31% | Included |
| Roundabout | Diameter | 107.19 | 357.30 | 35.73 | 3.57 | 159.51% | 15.95% | 1.60% | Excluded |
| Runway | Widest Point | 80.01 | 266.70 | 26.67 | 2.67 | 119.06% | 11.91% | 1.19% | Excluded |
| Sea Ice | No Dataset Presence | - | - | - | - | - | - | - | Excluded |
| Ship | Diagonal Length | 338.85 | 1129.50 | 112.95 | 11.30 | 504.24% | 50.42% | 5.04% | Excluded |
| Snowberg | No Dataset Presence | - | - | - | - | - | - | - | Included |
| Sparse Residential | Diagonal Length | 2450.38 | 8167.93 | 816.79 | 81.68 | 3646.40% | 364.64% | 36.46% | Merged |
| Stadium | Diagonal Length | 298.61 | 995.37 | 99.54 | 9.95 | 444.36% | 44.44% | 4.44% | Excluded |
| Storage Tank | Diagonal Length | 89.88 | 299.60 | 29.96 | 3.00 | 133.75% | 13.38% | 1.34% | Excluded |
| Tennis Court | Diagonal Length | 197.09 | 656.97 | 65.70 | 6.57 | 293.29% | 29.33% | 2.93% | Excluded |
| Terrace | Diagonal Length | 4445.10 | 14817.00 | 1481.70 | 148.17 | 6614.73% | 661.47% | 66.15% | Merged |
| Thermal/Power Station | Diagonal Length | 1050.45 | 3501.50 | 350.15 | 35.02 | 1563.17% | 156.32% | 15.63% | Excluded |
| Wetland | Diagonal Length | 8667.55 | 28891.83 | 2889.18 | 288.92 | 12898.14% | 1289.81% | 128.98% | Merged |

*Included as transition between land and water despite beach not necessarily being visible.

20% to testing. The split was performed randomly to prevent bias towards more representative class images. Ten RGB Landsat images, all captured in the summer of 2020, were acquired from the Planet Landsat image repository for dataset construction. These images were selected to include relevant land features, towns, and metropolitan areas. PlanetScope imagery was extracted but not incorporated at this phase of the study.

In order to determine which features should be included in an effective inter-GRD dataset, the authors quantitatively evaluated feature sizes visible from space at different GRDs. In considering the above datasets, the authors concluded that some smaller features, such as individual airplanes, sports fields, and intersections, would make up a very small part of a larger tile for even 3.0-meter GRDs, let alone 30.0-meter GRDs. Features which made up less than 35% of a 224-pixel square tile had their classes excluded from the inter-GRD dataset. Features were measured using tools within Google Earth Pro. At least three reasonably large examples of each feature were considered. The maximum feature size observed was used to promote fairness in inclusion. Measurements were made with respect to the widest point, diagonal length, or diameter depending upon the characteristics of the feature.

A depiction of the class-narrowing metrics can be referenced in Table II. Most features of obvious small size, such as individual buildings, sports fields, intersections, and overpasses, are excluded from the ideal inter-GRD dataset. While they may be relevant at 0.3-meter GRDs, they are barely visible at 3.0-meter GRDs, and not at all visible in 30.0-meter GRD imagery. More general geographic features, such as deserts, forests, islands, and lakes, are highly scalable between GRDs. For example, while a small lake may be visible at a low GRD and absent at a higher one, the class remains applicable at both. A large lake at a low GRD may instead be broken into tiles classified as coastline and sea, but all involved classes are still present and visible. Classes were further narrowed with the absence of sufficient representations for several classes in the Landsat imagery used, including chaparral, sea ice, and

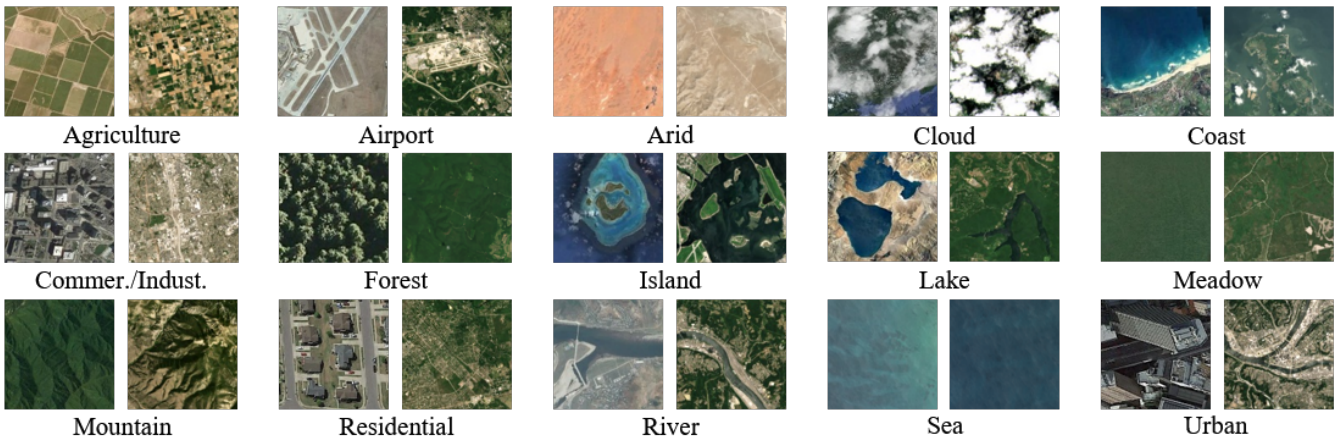


Fig. 3. NWPU-RESISC15 [21] (Left) and SHREC2020-30M15C (Right) Sample Images

snowberg.

Based on the results of the class refinement, a new variant of the NWPU-RESISC45 dataset was created to reduce the number of classes, promoting increased accuracy and reduced confusion. The urban class was handcrafted from urban images in dense residential, commercial, and industrial classes. Half of the commercial and industrial images were relegated to a combined commercial and industrial class. Half of the circular and rectangular farmland images were combined into an agriculture class. Non-urban dense residential and medium residential images were similarly merged with an even balance of 350 images maintained for each. The terrace class, while unmerged in the created NWPU-RESISC15 dataset permutation, was considered in the limited Landsat imagery in which it was present as either agriculture or meadow. Similarly, the wetland class was unmerged in the NWPU-RESISC15 dataset permutation, but considered coast or, in some rare cases, island in Landsat imagery occurrences. The sparse residential class fit the scale requirements, but seemed to provoke confusion with agricultural, forest, and other large geographic feature areas at large scales, thus sparse residential images are left out of the NWPU-RESISC15 dataset permutation. Several classes were renamed: desert to arid and beach to coast. The sea class was constructed from 700 256-pixel square sea tiles split from acquired Landsat imagery. A 700-image-per-class balance was maintained for this NWPU-RESISC15 dataset permutation.

Although the UC Merced and NWPU-RESISC45 datasets provide an adequate representation of imagery at GRDs as low as 0.3 meter per pixel, this study sought to construct a 30.0-meter GRD dataset for training and comparison. In pursuit of this goal, ten Landsat images were divided into 224-pixel square tiles meant to fit the input image size of the largest MobileNetV2 model. Each 36-megapixel, roughly $6300 \times 6500$-pixel, Landsat image generated 812 224-pixel square tiles. 3815 tiles across multiple images were sorted to represent the 15 classes consistently visible at a 30.0-meter GRD. This will be referred to as the SHREC2020-30M15C dataset. A sample image from each dataset of the 15 classes they share is included in Fig. 3 for reference. One may note an obvious difference of scales in some examples but a general feature similarity between each pair, especially for more natural geographic features.

### *B. Transfer Learning*

Model transfer learning was performed using TensorFlow 2.0 on ImageNet-pretrained MobileNetV2 models by Google AI. Google Colab was leveraged to create an accessible and tunable system for training a variety of models with different permutations. This system loaded the desired dataset and MobileNetV2 model, retrained a new classification head fit to the number of classes desired for 15 epochs, and then fine-tuned the lower layers of the base model for 15 additional epochs to improve feature extraction for remote-sensing imagery. Limited data augmentation was applied to the UC Merced and NWPU-RESISC15/45 datasets as horizontal and vertical flips in training. Further data augmentation was performed on the smaller of Landsat images, applying manual flips and 90-degree rotations similar to [22] to generate all eight possible geometric permutations of each input image in the training set. Trained models were tested on unseen data split from the original dataset before importation. Test data is never augmented or pulled from augmented training data to ensure fairness. Models were exported as .pb saved models and .tflite converted models. Within the scope of this submission, the largest (224-pixel, 1.00 depth multiplier) and smallest (96-pixel, 0.35 depth multiplier) MobileNetV2 model variants were trained.

### *C. Inter-GRD Model Testing*

To demonstrate the effectiveness of a model trained at a higher GRD on imagery of lower GRD, and vice versa, a form of cross-validation was conducted where each model was used to perform inference with the test data of another GRD examined. The results demonstrate the robustness of each model to varying GRDs. This study was limited to models focused on the same number and set of classes, so only the

NWPU-RESISC15 and SHREC2020-30M15C datasets could be tested at this time. The models were trained using the same procedure described above, but test data was swapped between them.

### D. Inference Benchmarking

Inference benchmarks for the tile-classification models were conducted on the platforms described in the background section and noted in Table I. Each platform served as an equivalent architecture to a targeted space computer system. A custom TensorFlow Lite 2.4 inference application was cross-compiled for the target CPU architectures. Inference performance was averaged over 100 runs for each model variant at each platform-supported thread count. Profiling mode was enabled, and three warmup runs were conducted for each inference set. The idle and inference power consumptions were measured for each platform. In addition, dynamic energy consumption per inference was calculated as the dynamic power multiplied by the execution time. Memory use for inference was determined for each platform by using the "free -s 1 –kilo" UNIX command, recording the idle and peak runtime memory use, and calculating the overall memory consumption for the app. All platforms were benchmarked starting from an idle state with only the app and any essential monitoring processes running.

## IV. Results

This section presents the results of this study. Inter-class confusion and fine-tuned model accuracy are evaluated. Performance benchmarks, including execution time, speedup, memory use, and dynamic energy consumption, are presented.

### A. Fine-Tuned Model Accuracy

Top-one accuracy results for the models trained on the UC Merced, NWPU-RESISC45, NWPU-RESISC15, and SHREC2020-30M15C datasets are included in Table III. UC Merced and NWPU-RESISC45 models are included as a baseline. UC Merced attains 96.19% model accuracy for 224-pixel tiles and 89.52% for 96-pixel tiles, likely due to clearer inter-class definition and fewer classes to contribute to confusion. NWPU-RESISC45 maintains 91.40% accuracy for 224-pixel tiles and 81.47% accuracy for 96-pixel tiles as well as a stable confusion matrix despite more than doubling the class count. The results reached in this study with existing datasets are competitive with the state of the art in [21] and [22].

The NWPU-RESISC15 permutation generated for this research demonstrates a 93.00% accuracy at 224-pixel tiles and an 87.57% accuracy on 96-pixel tiles. Compared to NWPU-RESISC45, this represents a 1.60% increase in overall accuracy for 224-pixel tiles and a 6.10% increase for 96-pixel tiles. The reduction of classes from 45 to 15 results in most classes meeting or exceeding 89% precision. The exceptions to this rule are the commercial/industrial and urban classes, the distinction of which is made difficult by the combination

TABLE III
Top-One Model Accuracy

| Dataset | Model Variant | | Accuracy Drop |
|---|---|---|---|
| | Depth 1.00 Size 224 | Depth 0.35 Size 96 | |
| UC Merced | 96.19% | 89.52% | 6.67% |
| NWPU-RESISC45 | 91.40% | 81.47% | 9.93% |
| NWPU-RESISC15 | 93.00% | 87.57% | 5.43% |
| SHREC2020-30M15C | 86.60% | 82.65% | 3.95% |

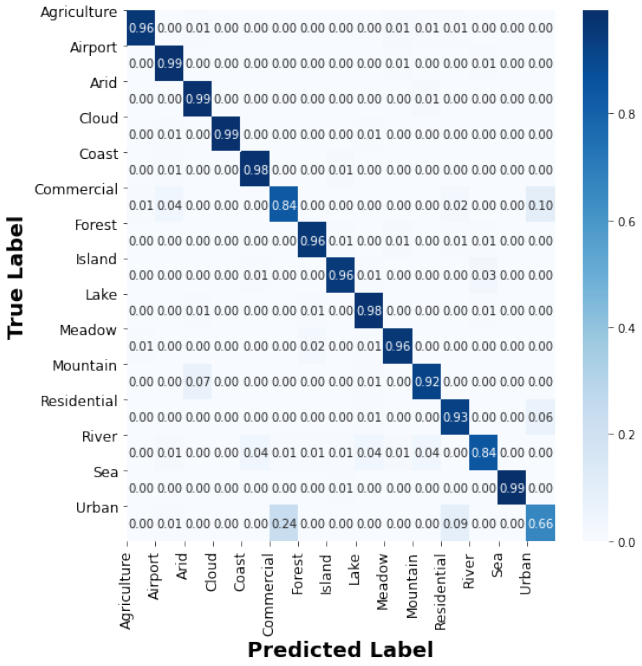


Fig. 4. Confusion Matrix for NWPU-RESISC15 Dataset Permutation

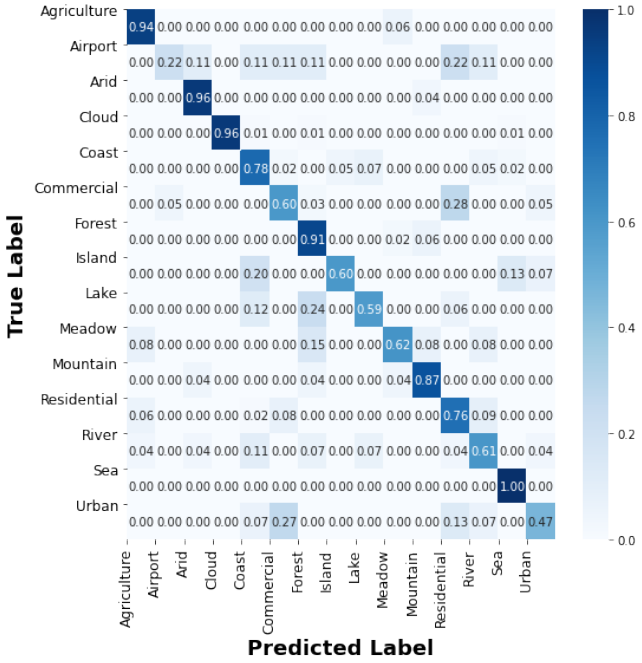


Fig. 5. Confusion Matrix for SHREC2020-30M15C Dataset

TABLE IV
PRECISION, RECALL, $F_1$ SCORE, AND SUPPORT

| Class | Dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | NWPU-RESISC15 | | | | SHREC2020-30M15C | | | |
| | Precision | Recall | $F_1$ Score | Support | Precision | Recall | $F_1$ Score | Support |
| Agriculture | 0.99 | 0.96 | 0.97 | 140 | 0.84 | 0.94 | 0.89 | 34 |
| Airport | 0.93 | 0.99 | 0.96 | 140 | 0.50 | 0.22 | 0.31 | 9 |
| Arid | 0.92 | 0.99 | 0.95 | 140 | 0.95 | 0.96 | 0.95 | 54 |
| Cloud | 1.00 | 0.99 | 0.99 | 140 | 1.00 | 0.96 | 0.98 | 79 |
| Coast | 0.96 | 0.98 | 0.97 | 140 | 0.73 | 0.78 | 0.75 | 41 |
| Commercial | 0.77 | 0.84 | 0.80 | 140 | 0.69 | 0.60 | 0.64 | 40 |
| Forest | 0.96 | 0.96 | 0.96 | 140 | 0.86 | 0.91 | 0.89 | 81 |
| Island | 0.96 | 0.96 | 0.96 | 140 | 0.82 | 0.60 | 0.69 | 15 |
| Lake | 0.93 | 0.98 | 0.95 | 140 | 0.67 | 0.59 | 0.62 | 17 |
| Meadow | 0.95 | 0.96 | 0.96 | 140 | 0.62 | 0.62 | 0.62 | 13 |
| Mountain | 0.94 | 0.92 | 0.93 | 140 | 0.71 | 0.87 | 0.78 | 23 |
| Residential | 0.88 | 0.93 | 0.90 | 140 | 0.75 | 0.76 | 0.75 | 66 |
| River | 0.94 | 0.84 | 0.89 | 140 | 0.61 | 0.61 | 0.61 | 28 |
| Sea | 1.00 | 0.99 | 1.00 | 140 | 0.98 | 1.00 | 0.99 | 246 |
| Urban | 0.80 | 0.66 | 0.73 | 140 | 0.64 | 0.47 | 0.54 | 15 |
| Accuracy | | | 0.93 | 2100 | | | 0.87 | 761 |
| Macro Average | 0.93 | 0.93 | 0.93 | 2100 | 0.76 | 0.73 | 0.73 | 761 |
| Weighted Average | 0.93 | 0.93 | 0.93 | 2100 | 0.86 | 0.87 | 0.86 | 761 |

method described earlier. The confusion matrix for this model is shown in Fig. 4 and discussed in Section V.

The SHREC2020-30M15C dataset attains a top-one accuracy of 86.60% for 224-pixel tiles and 82.65% for 96-pixel tiles. These results are higher than expected for its only 3815-image small-data approach and uneven dataset balance. The confusion matrix is visible in Fig. 5 and discussed in Section V. Table IV provides insight into the precision, recall, $F_1$ score, and number of support samples in the evaluation of the trained models. This especially helps to better illustrate the class imbalance problems encountered in the SHREC2020-30M15C dataset.

### B. Inter-GRD Study

With a wide variety of Earth-observation platforms capable of a variety of different GRDs, this study sought to investigate the amenability of models trained at larger GRDs to smaller ones, and vice versa. In this case, the low-GRD NWPU-RESISC15 dataset and high-GRD SHREC2020-30M15C dataset are compared. A comparison grid of test accuracy results is included in Table V.

Accuracies for the diagonals of each matrix are carried directly from previously noted results. Training on NWPU-RESISC15 is the most effective between GRDs, likely due to its varied scales, attaining 57.95% and 56.11% accuracy on SHREC2020-30M15C for large and small model variants, respectively. Even despite the much higher GRD of SHREC2020-30M15C, testing of its model on NWPU-RESISC15 still yields 39.33% and 34.43% accuracy for large and small model variants, respectively. Most interestingly, $F_1$ scores for more scale-invariant features remain high, as noted in Section V. Inter-GRD accuracy is quite consistent between large and small models, dropping by 4.90% for SHREC2020-30M15C and only 1.84% for NWPU-RESISC15.

TABLE V
INTER-GRD TRAINED MODEL ACCURACIES

| Large Models - Depth 1.00 Size 224 | | |
|---|---|---|
| | Tested Model | |
| Trained Model | NWPU-RESISC15 | SHREC2020-30M15C |
| NWPU-RESISC15 | 93.00% | 57.95% |
| SHREC2020-30M15C | 39.33% | 86.60% |
| **Small Models - Depth 0.35 Size 96** | | |
| | Tested Model | |
| Trained Model | NWPU-RESISC15 | SHREC2020-30M15C |
| NWPU-RESISC15 | 87.57% | 56.11% |
| SHREC2020-30M15C | 34.43% | 82.65% |

### C. Inference Benchmarks

This section presents benchmarks of execution time, speedup, memory use, and dynamic energy consumption. Execution-time benchmarks are included in Fig. 6. The authors note the log scale in the y-axis of the chart. These results focus on the largest and smallest MobileNetV2 variants. The best performing multithreaded runtimes for the largest model variants on each platform are 1600.13 ms for the A9, 198.27 ms for the dual A53, 214.10 ms for the single A53, and 97.60

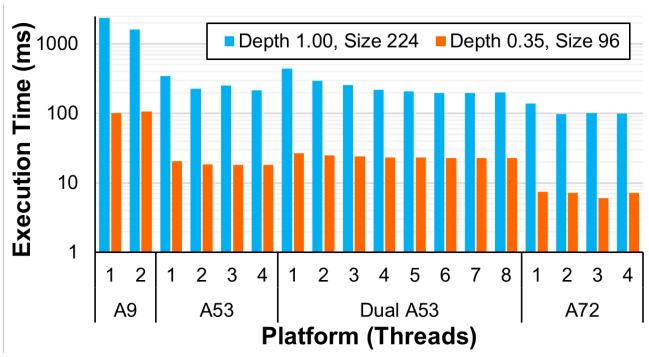


Fig. 6. MobileNetV2 Model Performance

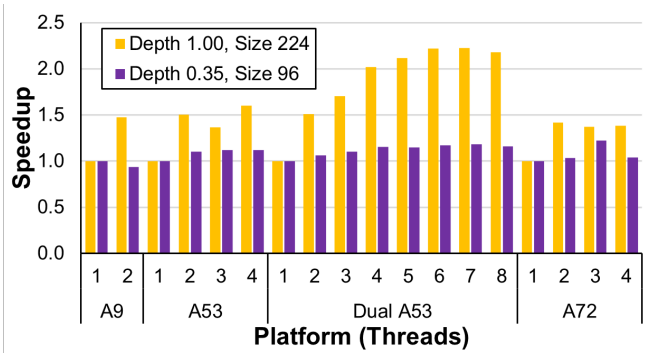


Fig. 7. MobileNetV2 Model Threading Speedup

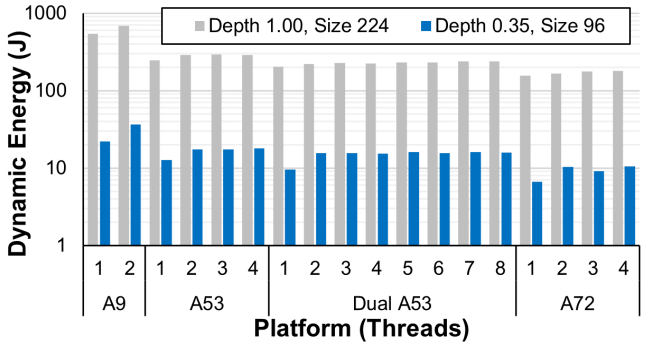


Fig. 9. MobileNetV2 Model Dynamic Energy Consumption

ms for the A72 per inference, respectively. In comparison, the smallest model variants yield ideal timings of 100.72 ms, 22.67 ms, 18.32 ms and 6.09 ms per inference, respectively. This amounts to performance gains of 15.89×, 8.75×, 11.68×, and 16.02×, respectively. Between architectures, there is an up to 8.07× gain from A9 to A53, an up to 3.00× gain from A53 to A72, and an up to 16.54× gain from A9 to A72. The tiling operation itself is considered negligible within the context of this study. Regardless of platform and even for very high-resolution images, the process of classifying the number of tiles generated requires a significantly larger period of time. Further, no preprocessing is required before inputting these tiles into the classifier.

The speedup at different thread counts for each platform is shown in Fig. 7. The large models see the only significant results, with the small-model runtimes too short to overcome thread-management overhead. The best-performing platform was the dual A53, peaking at 2.22× speedup, though it required seven threads to do so. The single A53 performed second best with 1.60× speedup on four threads. The A72, presumably due to architectural improvements, higher frequency, and increased memory bandwidth, improves speed enough that the large models have similar issues overcoming threading overhead.

Memory-use benchmarks for each platform are included in Fig. 8. For the CPUs tested, the peak memory use for the large models is 29.93 MB for A9, 25.10 MB for A53, and 20.94 MB for A72. The smaller models peak at 5.13 MB, 6.48 MB, and 5.78 MB, respectively. This represents a potential memory savings of up to 5.84×, up to 3.87×, and up to 3.82×, respectively. Memory use is not observed to correlate heavily with architectural improvement.

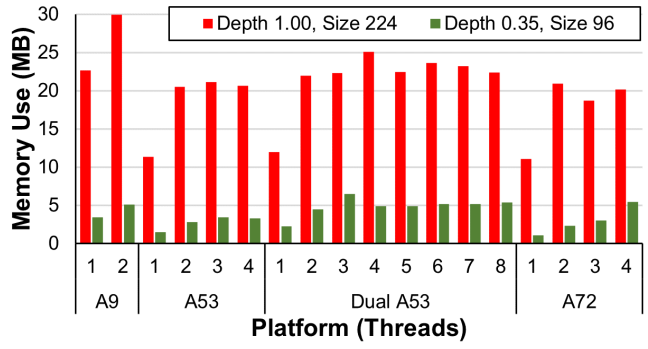


Fig. 8. MobileNetV2 Model Memory Use

Dynamic energy consumption results for the platforms tested are included as Fig. 9. The authors note the log scale in the y-axis of the chart. The lowest dynamic energy consumptions for large model inference are 542.4 J for the A9, 202.8 J for the A53, and 155.2 J for the A72. Smaller model sizes yield considerably lower minima of 22.2 J, 9.6 J, and 6.7 J, respectively. The transition from largest to smallest model variant thus represents energy savings of 95.91%, 95.25%, and 95.68%, respectively, averaging 94.07% reduction in dynamic energy consumption across all platforms and thread counts tested. Architectural improvements provide a drastic reduction in dynamic energy consumption, yielding improvements of up to 1.44× from A53 to A72, up to 2.67× from A9 to A53, and up to 3.49× from A9 to A72.

## V. Discussion

This section expands the discussion of trends from the results. More detailed inter-class accuracy insights are provided, especially for inter-GRD models. Implications of improved performance and dynamic energy consumption are explored.

### A. Model Accuracy

Closer inspection of inter-class accuracy from the NWPU-RESISC45 trained model confirms the authors' intuitions that classes based on smaller features are more difficult to accurately classify. The single-building church class attains only 80% accuracy. The unique palace class attains only 68% accuracy. There is also a high degree of confusion between the different densities of residential areas, with sparse, medium, and dense reaching 81%, 82%, and 85% accuracies, respectively. As discussed previously, many of the classes of the NWPU-RESISC45 dataset would not be visible at higher GRDs. The exclusion of smaller and easily confused classes in the NWPU-RESISC15 dataset shows improvement in accuracy as well as a refined list of classes that can be applied to a variety of GRDs. The confusion matrix of the NWPU-RESISC15 model represents a near-solid diagonal. The only

notable departure is some confusion between commercial and urban imagery, understandable given their similarity.

While the SHREC2020-30M15C dataset is unbalanced, the relative distribution of classes is consistent with a typical real-world distribution. Classes with reduced representation do demonstrate significantly lower $F_1$ scores, as low as 0.31 for the airport class and 0.61 for the river class. The confusion matrix shows notable inter-class confusion for this model between the commercial, residential, and urban classes, similar to NWPU-RESISC15. This lower accuracy is understandable given the parallel occurrence of commercial and residential, commercial and urban, residential and urban, or all three class features in a single tile. This classification problem transitions from simply distinguishing the presence of a class to selecting the best class representation given the amount and relationship of features present in the image tile. Many of the issues with this dataset are associated with the richness of features of different classes in the same tile. Many airports, for example, are located in densely commercial areas or near coastlines, two often mispredicted results for the airport class.

For all models observed, reducing the feature depth to 0.35 and the image size to 96 pixels impacts accuracy between 4-10%. This accuracy drop is considered acceptable given the goals of this study. The transition to smaller model variants is exceptionally beneficial when the performance and energy benchmarks are considered.

The inter-GRD study revealed that features that are more scale-invariant transfer best between models and achieve greater $F_1$ scores. For large-model training on NWPU-RESISC15 and testing on SHREC2020-30M15C, more geographic land-cover features see improved $F_1$ scores, including 0.75 for the arid class, 0.68 for the mountain class, and 0.37 for the agriculture class. For large-model training on SHREC2020-30M15C and testing on NWPU-RESISC15, while accuracy is considerably worse, similar positive trends in per-class accuracy exist, such as 0.55 for the river class, 0.44 for the lake class, and 0.33 for the island and mountain classes. The worst translating class for both combinations is urban at 0.00 and 0.06, respectively. The urban class's poor performance is attributed to discrepancies in scale between the two datasets, with even individual buildings present in the low-GRD NWPU-RESISC15 imagery, illustrated in Fig. 3. As the visible patterns for residential areas are more similar between the NWPU-RESISC15 and SHREC2020-30M15C datasets, it was hoped that some of the features for residential areas at different scales would transfer well. Unfortunately, the residential class performed poorly on both combinations, with 0.21 and 0.07 $F_1$ scores, respectively. It seems that the up to two order-of-magnitude transition from 0.3-meter NWPU-RESISC15 to 30-meter SHREC2020-30M15C imagery is too significant of a scale transition for these features to be transferable.

### *B. Inference Benchmarks*

It is important to recognize that many of the performance gains realized between platforms come from large differences in clock frequency. For example, compare the 650 MHz ARM Cortex-A9 and 1.5 GHz ARM Cortex-A72 platforms. The 32-bit A9 is another factor in reduced performance compared to the 64-bit A53 and A72. At the same time, each of these platforms aims to represent the architecture and frequency of a viable space computing system, justifying their selection. Tile classification provides opportunity for tuning tradeoffs in performance and situational awareness. For example, a high-resolution image may be tiled into 224-pixel square tiles. These tiles can then be input to the large model variants for maximum accuracy. Alternatively, the large tiles can be downsampled to as small as 96-pixel square tiles to increase throughput if faster decision-making is required with the tradeoff of reduced accuracy. As another option, the high-resolution image may be tiled to 96-pixel square for finer-grain classification with a modest speed increase. Assuming the ideal 16.02× speedup for small model variants, but with approximately 5.44× the number of smaller tiles extracted from the image, this will still amount to a 2.94× overall improvement while adding many more classification results.

Overall, the parallel performance, which measures the runtime of partitioning a single frame inference and processing all partitions in parallel, was quite poor in this study. This is likely due in no small part to the lightweight, low-latency nature of the MobileNetV2 model. The small data size and light model structure make it difficult to realize significant gains by dividing between threads. The models are memory bound, not compute bound. If one were to inspect a visual representation, the model itself is quite linear, with minimal task-level division. The lackluster gains from threading within a single inference operation led the authors to consider launching several single-threaded operations to batch process multiple images. This could improve parallel efficiency and increase throughput, allowing tiles to be classified more rapidly.

Memory use for the larger and more complex model variant seems to peak at two threads. The smaller model variant sees continued growth, albeit at a much slower rate, until roughly four threads. This is attributed to the larger relative size of per-thread memory when model size is small. Memory use is low for all platforms, making this approach highly amenable to memory-constrained systems.

The poor parallel performance observed in this study makes single-threaded operation the most energy efficient. However, this reinforces the possibility of using additional threads for batch processing of multiple images rather than model parallelism on a single image. Most interestingly, transitioning from the maximum image size and feature depth to the minimum results in an average 94.07% reduction in dynamic energy consumption across all platforms and thread counts. Mission-critical applications could see immense benefit from these gains. Especially for space platforms, dynamic energy consumption plays an important role. Consider classification on an interplanetary probe or rover platform where energy must be carefully managed. The opportunity to make 16.86× as many decisions may be more beneficial than attaining the highest possible accuracy.

## VI. Future Research

There are several directions for further research stemming from this study. Toward GRD-agnostic models, there remains a dilemma of accurately classifying relevant features in high-GRD imagery due to the large area represented and thus the high number of varied features present. Many high-GRD sample images might contain, for example, urban, forest, and river components in a single image tile. It is thus difficult for a single-decision classifier to extract and interpret the most relevant features. Several options, including multi-label classification, hierarchical classifiers, and multi-head classification, with each classification head focusing on one aspect of the image contents, are being considered as a solution in an extension of this research. Research is underway into a multi-head classification approach in which land cover, more geographic features like rivers and coastlines, and land use, primarily human-constructed features, would be handled by separate classification heads, each more effectively tuned to isolate target features. This approach aims to improve the effectiveness of classifying feature-dense high-GRD imagery.

It would be helpful to have a larger set of datasets for which to investigate inter-GRD relationships, perhaps including GRD increments of 3.0 meters, 6.0 meters, 9.0 meters, and so on, per pixel. This is a primary target for further extension of this study. This research also creates an opportunity for improved applications of planning, rule-based systems, and autonomy based on tile classes. Providing a system with knowledge of the classes of many components of an image and exploring the relationship between them presents an effective means for space situational awareness. Exploring self-supervised learning will be beneficial to reduce labelling burden for Earth-observation imagery but will need to be approached carefully given the feature richness of high-GRD imagery. The full design space of model image sizes and feature depths could be examined with further study, especially with regard to optimizing accuracy for dynamic energy consumption. Additional platforms and models could be considered and benchmarked in additional research. Given the poor parallel performance, future efforts aim to use threads to batch process image tiles, which should significantly improve gains. Development constraints on several chosen platforms prevented deep-learning model quantization performance improvements from being competitively considered. This will be further explored in a future extension of this research.

## VII. Conclusions

This research occurs at the convergence of high-resolution and low-GRD sensing, capable onboard computing, and minimal-parameter, optimized DCNNs for computer-vision tasks. This study demonstrated MobileNetV2 models fine-tuned on new and existing Earth-observation datasets, including training the new SHREC2020-30M15C dataset to 86.60% accuracy despite its high GRD and the associated inter-class feature richness. The authors quantitatively refined classes by visible feature size in pursuit of improved GRD-agnostic models. An inter-GRD error study quantifies model effectiveness when applied at different GRDs and depicts exceptional results for scale-invariant geographic features but poorer handling of human-constructed land development. Execution time, parallel performance, memory use, and dynamic energy consumption benchmarks on several CPU architectures with flight heritage or mission applicability confirm onboard processing amenability for these models. Tile classification demonstrates high versatility in that running smaller image-size and shallower feature-depth model variants can yield an 8-16× performance improvement and 94.07% average reduction in dynamic energy consumption with an only 4-10% reduction in accuracy. This tile-classification system can provide a versatile and high-performance solution for growing Earth-observation demand.

## Acknowledgment

This research was supported by the NSF Center for Space, High-performance, and Resilient Computing (SHREC) industry and agency members and by the IUCRC Program of the National Science Foundation under Grant No. CNS-1738783. The authors would like to thank and acknowledge several key contributors to the success of this project. David Langerman provided critical assistance and feedback on dataset development, model training, and results analysis. Seth Roffé, Tyler Garrett, Michael Cannizzaro, Clara Ferreira, Kyle Buettner, Antony Gillette, Collin Abidi, and Sebastian Sabogal provided helpful feedback and results review. Dr. Andrew Pineda and Evan Kain at the Air Force Research Laboratory Space Vehicles Directorate provided insights and feedback on preliminary versions of this research. Planet provided research access to Landsat and PlanetScope imagery used for training the models in this study.

## References

[1] BAE Systems, "RAD6000 Space Computers," Tech. Rep., 2006.

[2] ——, "RAD750® Radiation-Hardened PowerPC Microprocessor," Tech. Rep., 2008.

[3] ——, "RAD5545™ multi-core system-on-chip Power Architecture processor," 3 2017.

[4] C. Wilson and A. George, "CSP hybrid space computing," in *Journal of Aerospace Information Systems*, vol. 15, no. 4, 2018.

[5] A. D. George and C. M. Wilson, "Onboard Processing with Hybrid and Reconfigurable Computing on Small Satellites," *Proceedings of the IEEE*, vol. 106, no. 3, 2018.

[6] N. Buonaiuto, C. Kief, M. Louie, J. Aarestad, B. Zufelt, R. Mital, D. Mateik, R. Sivilli, and A. Bhopale, "Satellite Identification Imaging for Small Satellites Using NVIDIA," *Small Satellite Conference*, 2017.

[7] J. MacKinnon, T. Ames, D. Mandl, C. Ichoku, L. Ellison, J. Manning, and B. Sosis, "Classification of Wildfires from MODIS Data Using Neural Networks," NASA Goddard Space Flight Center, Tech. Rep., 2017.

[8] J. Manning, D. Langerman, B. Ramesh, E. Gretok, C. Wilson, A. D. George, J. Mackinnon, and G. Crum, "Machine-Learning Space Applications on SmallSat Platforms with TensorFlow," *32nd Annual AIAA/USU Conference on Small Satellites*, 2018.

[9] A. P. Arechiga, A. J. Michaels, and J. T. Black, "Onboard Image Processing for Small Satellites," in *Proceedings of the IEEE National Aerospace Electronics Conference, NAECON*, vol. 2018-July, 2018.

[10] T. A. Estlin, B. J. Bornstein, D. M. Gaines, R. C. Anderson, D. R. Thompson, M. Burl, R. Castaño, and M. Judd, "AEGIS automated science targeting for the MER Opportunity rover," 2012.

[11] M. Ono, T. J. Fuchs, A. Steffy, M. Maimone, and J. Yen, "Risk-aware planetary rover operation: Autonomous terrain classification and path planning," in *IEEE Aerospace Conference Proceedings*, vol. 2015-June, 2015.

[12] K. Otsu, M. Ono, T. J. Fuchs, I. Baldwin, and T. Kubota, "Autonomous Terrain Classification with Co-and Self-Training Approach," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, 2016.

[13] B. Rothrock, J. Papon, R. Kennedy, M. Ono, M. Heverly, and C. Cunningham, "SPOC: Deep learning-based terrain classification for Mars rover missions," in *AIAA Space and Astronautics Forum and Exposition, SPACE 2016*, 2016.

[14] D. Qiu, B. Rothrock, T. Islam, A. K. Didier, V. Z. Sun, C. A. Mattmann, and M. Ono, "SCOTI: Science Captioning of Terrain Images for data prioritization and local image search," *Planetary and Space Science*, vol. 188, 2020.

[15] X. Chen, S. Xiang, C. L. Liu, and C. H. Pan, "Vehicle detection in satellite images by parallel deep convolutional neural networks," in *Proceedings - 2nd IAPR Asian Conference on Pattern Recognition, ACPR 2013*, 2013.

[16] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2015-October, 2015.

[17] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks," 2 2015.

[18] P. Lassalle, J. Inglada, J. Michel, M. Grizonnet, and J. Malik, "A Scalable Tile-Based Framework for Region-Merging Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 10, 2015.

[19] Q. Zhu, Y. Zhong, B. Zhao, G. S. Xia, and L. Zhang, "Bag-of-Visual-Words Scene Classifier with Local and Global Features for High Spatial Resolution Remote Sensing Imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, 2016.

[20] H. Alhichri, E. Othman, M. Zuair, N. Ammour, and Y. Bazi, "Tile-based semisupervised classification of large-scale VHR remote sensing images," *Journal of Sensors*, vol. 2018, 2018.

[21] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," 2017.

[22] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," *Sensors (Switzerland)*, vol. 20, no. 7, 2020.

[23] P. B. de Selding, "For DigitalGlobe, Government Business Steady But Commercial Disappoints," *SpaceNews*, 10 2015.

[24] D. P. Roy, M. A. Wulder, T. R. Loveland, W. C.E., R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy, T. A. Scambos, C. B. Schaaf, J. R. Schott, Y. Sheng, E. F. Vermote, A. S. Belward, R. Bindschadler, W. B. Cohen, F. Gao, J. D. Hipple, P. Hostert, J. Huntington, C. O. Justice, A. Kilic, V. Kovalskyy, Z. P. Lee, L. Lymburner, J. G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R. H. Wynne, and Z. Zhu, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sensing of Environment*, vol. 145, 2014.

[25] J. G. Masek, M. A. Wulder, B. Markham, J. McCorkel, C. J. Crawford, J. Storey, and D. T. Jenstrom, "Landsat 9: Empowering open science and applications through continuity," *Remote Sensing of Environment*, vol. 248, 2020.

[26] J. Pushparaj and A. V. Hegde, "A comparative study on extraction of buildings from Quickbird-2 satellite imagery with & without fusion," *Cogent Engineering*, vol. 4, no. 1, 2017.

[27] DigitalGlobe, "WorldView-1 Data Sheet," Tech. Rep., 2019.

[28] Gunter's Space Page, "WorldView 2,3," Tech. Rep., 2021.

[29] DigitalGlobe, "WorldView-4 Data Sheet," Tech. Rep., 2019.

[30] S. Sabogal, P. Gauvin, B. Shea, D. Sabogal, A. Gillette, C. Wilson, A. Barchowsky, A. D. George, G. Crum, and T. Flatley, "SSIVP : Spacecraft Supercomputing Experiment for STP-H6," *31th Annual AIAA/USU Conference on Small Satellites*, 2017.

[31] S. Roffe, T. Schwarz, T. Cook, N. Perryman, J. Goodwill, E. Gretok, A. Phillips, M. Moran, T. Garrett, and A. George, "CASPR: Autonomous Sensor Processing Experiment for STP-H7," in *2020 AIAA Small Satellite Conference*, Logan, UT, 8 2020.

[32] Planet, "Planet at a Glance: See Change, Change the World," Tech. Rep., 2019.

[33] ——, "PlanetScope Solutions Overview," Tech. Rep., 2020.

[34] ——, "SkySat Solutions Overview," Tech. Rep., 2020.

[35] T. Zheng, M. H. Bergin, S. Hu, J. Miller, and D. E. Carlson, "Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach," *Atmospheric Environment*, vol. 230, 2020.

[36] J. Park, J. Lee, K. Seto, T. Hochberg, B. A. Wong, N. A. Miller, K. Takasaki, H. Kubota, Y. Oozeki, S. Doshi, M. Midzik, Q. Hanich, B. Sullivan, P. Woods, and D. A. Kroodsma, "Illuminating dark fishing fleets in North Korea," *Science Advances*, vol. 6, no. 30, 2020.

[37] SpaceNet LLC, "SN7: Multi-Temporal Urban Development Challenge," Tech. Rep., 2020.

[38] Planet, "Planet Partners With SpaceNet For Multi-Temporal Urban Development Challenge," Tech. Rep., 2020.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, 2016.

[41] F. Iandola and K. Keutzer, "Keynote: Small neural nets are beautiful: Enabling embedded systems with small deep-neural-network architectures," in *2017 International Conference on Hardware/Software Codesign and System Synthesis, CODES+ISSS 2017*, 2017.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[44] A. Howard, M. Sandler, B. Chen, W. Wang, L. C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, Q. Le, and H. Adam, "Searching for mobileNetV3," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, 2019.

[45] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[46] B. Wei, X. Shen, and Y. Yuan, "Remote Sensing Scene Classification Based on Improved GhostNet," in *Journal of Physics: Conference Series*, vol. 1621, no. 1, 2020.

[47] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2010.

[48] "PYNQ-Z2 Reference Manual v1.0," Tech. Rep., 5 2018.

[49] X. Iturbe, D. Keymeulen, P. Yiu, D. Berisford, R. Carlson, K. Hand, and E. Ozer, "On the use of system-on-chip technology in next-generation instruments avionics for space exploration," in *IFIP Advances in Information and Communication Technology*, vol. 483, 2016.

[50] S. G. Kanekal, L. Blum, E. R. Christian, G. Crum, M. Desai, J. Dumonthier, A. Evans, A. D. Greeley, S. Guerro, S. Livi, K. LLera, J. Lucas, J. MacKinnon, J. Mukherjee, K. Ogasawara, N. Paschalidis, D. Patel, E. Pollack, S. Riall, Q. Schiller, G. Suarez, and E. J. Summerlin, "The MERiT Onboard the CeREs: A Novel Instrument to Study Energetic Particles in the Earth's Radiation Belts," *Journal of Geophysical Research: Space Physics*, vol. 124, no. 7, 2019.

[51] R. Doyle, R. Some, G. Mounce, S. Horan, and M. Lowry, "High Performance Spaceflight Computing (HPSC) Next--Generation Space Processor (NGSP) A Joint Investment of NASA and AFRL," Tech. Rep., 2014.

[52] Linaro Ltd., "HiKey LeMaker Hardware User Manual," Tech. Rep., 2021.

[53] Hardkernel Ltd., "ODROID-C2 User Manual," Tech. Rep., 11 2016.

[54] NXP Semiconductors, "NXP i.MX 8QuadMax MIMX8QM6AVUFFAB," Tech. Rep., 2021.

[55] Raspberry Pi (Trading) Ltd., "Raspberry Pi 4 Model B Datasheet Release 1," Tech. Rep., 6 2019.