Wake Vision: A Large-scale, Diverse Dataset and Benchmark Suite for TinyML Person Detection

Colby Banbury*[1], Emil Njor*[1,2], Matthew Stewart[1], Pete Warden[3],
Manjunath Kudlur[3], Nat Jeffries[3], Xenofon Fafoutis[2], and Vijay Janapa Reddi[1]

[1] Harvard University
[2] Technical University of Denmark
[3] Useful Sensors

wakevision.ai

Abstract. Machine learning applications on extremely low-power devices, commonly referred to as tiny machine learning (TinyML), promises a smarter and more connected world. However, the advancement of current TinyML research is hindered by the limited size and quality of pertinent datasets. To address this challenge, we introduce Wake Vision, a large-scale, diverse dataset tailored for person detection—the canonical task for TinyML visual sensing. Wake Vision comprises over 6 million images, which is a hundredfold increase compared to the previous standard, and has undergone thorough quality filtering. Using Wake Vision for training results in a 2.41% increase in accuracy compared to the established benchmark. Alongside the dataset, we provide a collection of five detailed benchmark sets that assess model performance on specific segments of the test data, such as varying lighting conditions, distances from the camera, and demographic characteristics of subjects. These novel fine-grained benchmarks facilitate the evaluation of model quality in challenging real-world scenarios that are often ignored when focusing solely on overall accuracy. Through an evaluation of a MobileNetV2 TinyML model on the benchmarks, we show that the input resolution plays a more crucial role than the model width in detecting distant subjects and that the impact of quantization on model robustness is minimal, thanks to the dataset quality. These findings underscore the importance of a detailed evaluation to identify essential factors for model development. The dataset, benchmark suite, code, and models are publicly available under the CC-BY 4.0 license, enabling their use for commercial use cases.

Keywords: Dataset · Person Detection · TinyML · Benchmarks

1 Introduction

To reduce the energy consumption and cost of deploying Machine Learning (ML) deployments, researchers have pioneered a new field of ultra-low-power machine learning (TinyML) [1, 3, 8, 35]. TinyML aims to co-locate ML models with the

---

* These authors contributed equally to this work

sensors by deploying the model onto always-on microcontrollers (MCUs) and TinyML accelerators, thereby reducing the energy cost of transmitting data or constantly running larger processors. Deploying models on these tiny devices imposes severe constraints, typically confining the model's size to a few hundred kilobytes, nearly four orders of magnitude less memory than mobile phones [2].

The field of TinyML requires large-scale datasets to facilitate high-quality research. However, the extreme constraints imposed by the hardware necessitate that TinyML models be compact and efficient, rendering them unsuitable for traditional, complex ML tasks. As a result, conventional ML datasets like ImageNet [9], designed for more resource-intensive models, are often ill-suited for TinyML research endeavors. While existing TinyML datasets [7, 34] have been valuable resources, they tend to be limited in scale, making it challenging to train production-grade TinyML models solely on these datasets.

To enable and mature the field of TinyML research, we present the Wake Vision dataset, a large dataset for person detection, the quintessential vision use case for TinyML [1]. Person detection is a binary image classification task where a model detects whether a person is present. While simple, this task has many important use cases, such as occupancy detection [29], smart HVAC/lighting [36], or acting as an always-on 'wake model' in a larger ML system [14].

Wake Vision has over 6M images, ~100x more than the prior state-of-the-art person detection dataset, Visual Wake Words (VWW) [7]. Wake Vision is derived from the Open Images v7 dataset [19, 21] and is permissively licensed. Wake Vision demonstrates a 2.41% accuracy improvement over VWW.

Due to the limited capacity of TinyML models, data quality is critical [26], however techniques, such as knowledge distillation [15], can leverage large, noisy datasets more effectively. For this reason, we choose to release two training sets: one that prioritizes dataset size, Wake Vision (Large), and one that prioritizes data quality, Wake Vision (Quality). We find that Wake Vision (Quality) out performs Wake Vision (Large), but we release both training sets to enable research into the tradeoffs between quantity and quality of data for TinyML.

Person detection systems are deployed, and expected to work, in challenging settings, such as low lighting. However, because these challenging settings are uncommon in images sourced from the internet, high-level metrics, like test set accuracy, may overestimate a model's real world performance. To address this issue we introduce a suite of five fine-grain benchmarks, which contain only images of a particular demographic or scenario. Our benchmarks cover distance, lighting, depictions, and perceived gender and age. Finally, we use our benchmark suite to conduct a case study on the performance effects of typical TinyML compression. This analysis demonstrates the usefulness of our benchmark and can inform future TinyML model design.

In summary, we present the following novel contributions:

– A new person detection dataset, called Wake Vision, which is an easily accessible, permissively licensed dataset that is ~100x larger than the prior standard.

- A suite of five fine-grain benchmark sets that evaluate a model's fairness and robustness in challenging settings.
- A case study on the effect of TinyML scaling the highlights the importance of fine grain analysis.

2 Related Work

The current de facto standard dataset for person detection is the VWW dataset [1, 2, 7, 8, 23]. To our knowledge, this is the only current open-source dataset explicitly meant for person detection. Unfortunately, the VWW dataset is small and cannot be directly downloaded. Instead, it must be regenerated from MS-COCO [24], thus limiting its accessibility and usefulness.
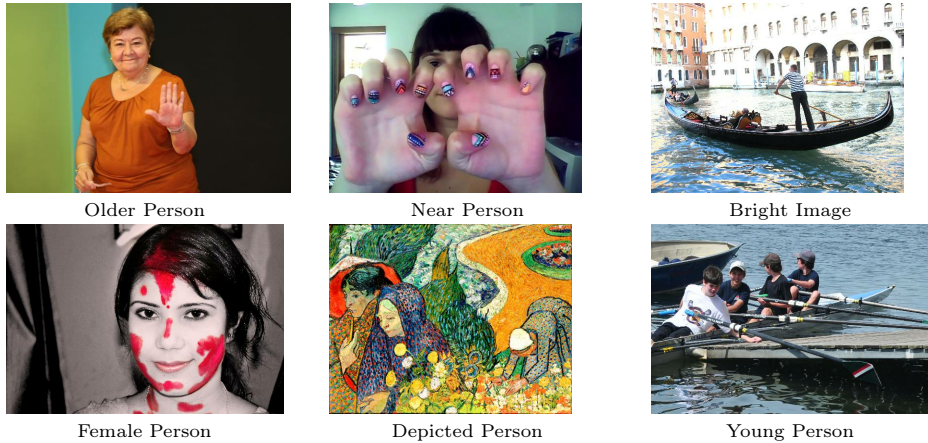
Apart from dedicated person detection datasets, some general image classification datasets contain a person label, which makes it possible to create person detection models based on them. Examples include the Cifar-100 dataset [20], and the PASCAL Visual Object Classes dataset [11]. However, using general image classification datasets can lead to poor perceived performance of TinyML models, as they do not adequately represent the open "no-person" class [7].

Tab. 1 shows that our Wake Vision dataset contains up to almost two orders of magnitude more images than any other public and permissively licensed dataset. Furthermore, Wake Vision is the only person detection dataset to provide a fine-grained benchmark suite and be officially distributed to popular dataset services such as TensorFlow Datasets (TFDS) and Hugging Face Datasets.

Furthermore, Wake Vision integrates fine-grained benchmarks, which are a novel addition to existing vision data sets for TinyML. These fine-grained benchmark sets, which include distance, lighting, depictions, and perceived gender and age variations, allow for a deeper investigation into the effects of input and model compression on specific demographic or scenario subsets. Several research papers report little performance degradation from model compression on top-line metrics such as top-1 test set accuracy [6,10,13,18]. However, recent research suggests that this focus on top-line metrics can hide poor performance on critical subsets of a dataset, *e.g.*, subsets relating to sensitive information and fairness [16, 17].

Table 1: A comparison of person detection and image classification datasets.

Dataset	Total Images	# of Person Images			Fine-Grain Filtering	Suitable for TinyML
		Train	Validation	Test		
Wake Vision (Quality)	**1,423,690**	**675,411**	**9,106**	**27,328**	✓	✓
Wake Vision (Large)	**6,477,906**	**3,238,953**	-	-	✗	✓
Visual Wake Words [7]	123,287	36,000	3,926	19,107	✗	✓
CIFAR-100 [20]	60,000	2,500	-	500	✗	✗
PASCAL VOC 2012 [11]	11,530	1,994	2,093	-	✗	✗

| Older Person | Near Person | Bright Image |
| Female Person | Depicted Person | Young Person |

Fig. 1: Sample images from the Wake Vision fine-grained benchmark suite.

3   Wake Vision Dataset

We introduce the Wake Vision dataset, the largest dataset for TinyML person detection. The dataset's size, quality, and detailed metadata enable new avenues of TinyML research. Wake Vision is two orders of magnitude larger than prior comparable datasets and can be adapted based on the target use case. Fig. 1 illustrates examples of images in the dataset. The dataset is permissively licensed for research and commercial use. Images are sourced from Flickr [12] images listed as having a CC-BY 2.0 license Labels are derived from Open Images and licensed under the CC-BY 4.0 license. In this section, we discuss our methodologies for label generation, data filtering, and error correction, as well as discussing the usability of the dataset and directly comparing it to the Visual Wake Dataset.

3.1   Label Generation

While a sufficiently large person detection dataset is indispensable for TinyML research, it can be challenging to obtain. Manually labelling millions of images would be prohibitively expensive for a nascent research field like TinyML.

Therefore, we leverage existing large-scale data efforts to to generate Wake Vision automatically. The base label in Wake Vision is a binary person/non-person label. We derive these labels from Open Images, which contain both image-level and bounding box labels. Image-level labels describe which objects are present in an image but not localized to a specific point in the image.

An image can contain several image-level labels, where as bounding box labels provide a set of four coordinates forming a box around the object. The bounding box label classes of Open Images are hierarchically structured so that one class can be a subcategory or a part of another class, *e.g.*, a "Woman" is a subcategory of a "Person," and a "Human Hand" is a part of "Person."

We observe that the bounding box labels are generally more accurate and provide more information for data filtering but are less numerous. The Open Images training set has over 3 million images with image-level person labels, but only ~ 700,000 images with a person bounding box. The Open Images validation and test sets are fully labeled with image-level and bounding box labels. Consequently, this presents a trade-off between more data (image-level labels) or higher-quality labels (bounding box labels), as different use cases may have different requirements for the training data (*e.g.*, active learning).

We provide two Wake Vision training sets: Wake Vision (Large) and Wake Vision (Quality). One larger set, labeled via Open Images image-level labels, and one smaller set, labeled via Open Images bounding box labels (Tab. 1). In our experiments, the Wake Vision (Quality) training set produce more accurate models than the Wake Vision (Large) training set, strongly suggesting that label quality is more important than quantity in our setting.

Large Scale Training Set from Image-Level Labels: Image-level labels in Open Images contain little additional information in addition to the labels. The only additionally relevant information is a confidence property, which represents how certain it is that a label is correct. The confidence ranges from 0-1 for labels generated by machines, and is either 0 or 1 for labels negatively or positively verified by humans. We use this property to exclude images from Wake Vision if they contain only low confidence person labels.

High Quality Training Set From Bounding Box Labels: Bounding Box labels in Open Images, in contrast to image-level labels, are all verified and localized by humans. This ensures that bounding box labels are less likely to be false positives. Bounding box labels can also be used to calculate an approximation of the area taken up by a person. We use this as a proxy for the distance of a person in the image, and exclude images where all person labels are far away, *i.e.*, take up only a small portion of the image. An example of an image with a far away person can be found in Fig. 2a. Finally, bounding box attributes include a depiction flag, which represents whether a label is a depiction. See an example of an image containing a person with the depiction flag set in Fig. 2b. By default we use this flag to make sure that we do not consider a depiction a person.

We allow users to adapt the Wake Vision dataset to meet the needs of specific usecases by adding several configurable options in the open-source dataset creation script. A comprehensive list of Open Images labels related to Wake Vision can be found in Appendix I.

Training Set Comparison To compare the performance of the two training dataset sets, we train a MobileNetV2-0.25 [31] model on each dataset for an equal number of steps and compare its performance on the common test set (Tab. 2). Given the amount of label errors in the Wake Vision test set shown in Tab. 4, a perfect person detection model would achieve an accuracy of ~93%.

(a) An example of a person far away

(b) An example of a depicted person
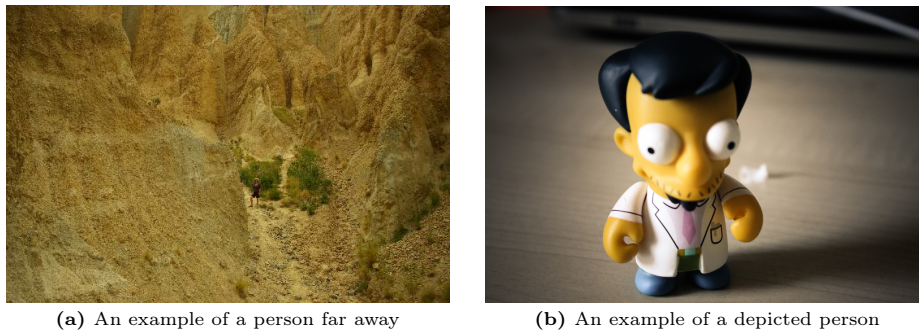
Fig. 2: Examples of challenging outlier images.

The filtered training set generated from the bounding box labels outperforms the larger training set generated from the image-level labels by 2.65% test accuracy. This indicates that label quality is more important than quantity in this setting. Interestingly, although the gap between the two is reduced to just 0.36% when training exclusively on soft labels from a teacher model, the bonding box training set is still superior. This indicates that the difference in quality extends beyond just the labels and that our data filtering leads to an improvement on it's own.

Table 2: Accuracy on Wake Vision Test set of a MobileNetV2-0.25 model trained on image-level and bounding box labels, with and without distillation. A MobileNetV2-1.0 model trained on the bonding box set is used as the teacher in distillation.

Training set	Label Source	Dataset Size	Accuracy	Distilled Accuracy
Wake Vision (Large)	Image-Level	6,477,906	86.00%	88.96%
Wake Vision (Quality)	Bounding Box	1,350,822	88.65%	89.32%

Despite the worse performance in this setting, the larger image-level training set may be useful for training higher capacity models for longer or when doing active learning [30] or distillation [5]. As our distillation results indicate, further data-centric filtering techniques may be needed to fully leverage the image-level training set. We experimented with tripling the model size and doubling the training time, but the difference in accuracy remained relatively consistent.

The two training sets provide a useful foundation for future work on the importance of data quality vs. quantity. Wake Vision can enable further research on data-centric TinyML [25, 26], specifically uncovering the relationship between model capacity and the sensitivity to training data quality.

3.2 Error Correction

Labeling errors are a challenge in computer vision [4, 28] that limit progress and obscures true performance. This is especially prevalent in large datasets

Table 3: Amount of label errors identified and corrected using Confident Learning.

Dataset Split	Total Size	Suggested Errors	Corrected Errors	Suggestion Accuracy
Validation	16,964	632	81	12.82%
Test	51,282	1672	267	15.97%

that use complex labeling pipelines to scale while managing costs. Recognizing the importance of label quality, we take measures to estimate and improve the accuracy of its labels, aligning with best practices in the field of computer vision.

Wake Vision's labels are derived from the Open Images labels; therefore, the errors that occur in the Open Images labeling pipeline are inherited by Wake Vision. All labels in the Open Images dataset are originally machine-generated candidate labels. Many of these labels are later verified by human annotators, and a subset of them are given bounding boxes and additional annotations.

While the machine-generated label phase aims to identify objects efficiently, any instances missed during this initial step are unlikely to be captured in downstream phases. Consequently, the Open Images dataset may contain numerous false negative labels, referring to images where an object is present, but lacks the corresponding label. Therefore, additional measures are necessary to identify and correct such labeling omissions to ensure the dataset's integrity.

Given the substantial scale of Wake Vision, manually correcting label errors across the entire dataset becomes an arduous undertaking. To address this challenge, we employ the Confident Learning technique [27] to intelligently identify potential label errors. We selected Confident Learning as prior work has demonstrated its capability to find label errors in large datasets [28]. Once these suspected mislabeled instances are flagged, we methodically inspect and correct them through a manual verification process.

As shown in Tab. 3, the confident learning process identified a large amount of possible label errors in Wake Vision's Validation and Test sets. However, only between 12 and 16% of these possible label errors were legitimate errors. We correct these identified label errors in the Wake Vision validation and test sets.

Given this low acceptance rate of label issues identified by Confident Learning, we conclude that we cannot automate label cleaning to a point where no human in the loop is needed. Based on that we also conclude that this strategy too human-intensive to be applied to the much larger training sets.

We report our estimated label error rate after label corrections versus VWW in Tab. 4. We find that after our corrections, the VWW dataset has a top-level error rate of 7.8%, which is comparable to Wake Vision's top-level error rate of 6.8%. Notably, about a third of the Wake Vision label errors stem from depictions missing the depiction flag. As we by default do not consider depictions as persons, missing this flag results in a label error. Not considering the depiction label errors, Wake Vision has closer to half the number of errors of VWW.

Table 4: Percentage (and count) of label errors in a 500 sample subset of the VWW and Wake Vision test sets. Percent of label errors caused by misclassified depictions are reported.

Dataset	Label Error Rate	% Errors From Depictions
Visual Wake Words	7.8% (39)	0% (0)
Wake Vision	6.8% (34)	38.2% (13)

VWW does not, to the authors knowledge, include depictions. Thus no depiction errors are present for the VWW dataset. In real world settings, depictions are everywhere, therefore we expect that Wake Vision will prove to be a more realistic dataset. ML systems based on Wake Vision will therefore likely be able to handle real world scenarios better than equivalent systems based on VWW.

We are in the process of crowd-sourcing label corrections for the validation and test sets to eliminate the remaining errors. However, despite these efforts, Wake Vision, in its current state, already offers a significant improvement over the VWW dataset, as detailed in Section 3.4. Although further enhancements to label quality are underway, Wake Vision presents a valuable and directly improved resource compared to its predecessor VWW dataset.

3.3   Usability

Wake Vision is available through TensorFlow Datasets [33] and HuggingFace Datasets [22] to be easily accessed and used by the community. The images and labels are rehosted to ensure the dataset will not be changed over time due to dead links, making it a more viable benchmark. The rehosted labels are generated according to our default dataset configuration, further described in Appendix F.

3.4   Comparison to Visual Wake Words

Wake Vision serves as a drop-in improvement over VWW, thanks to its size and our extensive filtering. To compare the benefits of Wake Vision against VWW, we train two identical models using the same recipe for the same number of steps, one utilizing VWW's training set and the other employing Wake Vision's training set. For this comparative evaluation, both trained models are MobileNetV2 architectures with a width modifier of 0.25, which have previously served as benchmark TinyML models [7]. After training these identical models using the respective datasets' training sets with the same recipe for an equal number of steps, we cross-evaluate their performance on the corresponding test sets.

In Tab. 5 we demonstrate a direct 0.5% improvement on VWW's own test set, which illustrates that Wake Vision is a direct improvement over VWW, not simply a domain shift. We additionally achieve a 2.41% improvement over VWW on Wake Vision's Test set, indicating that the new test set is more challenging.

Table 5: Accuracy on the Wake Vision and VWW test sets by models trained on each dataset's respective training set.

		Train	
		VWW	Wake Vision
Test	VWW	88.00%	**88.50%**
	Wake Vision	86.27%	**88.68%**

4 Fine Grain Benchmark Suite

Test sets generated directly from images posted to the internet are not reflective of real world use cases as the distribution is biased towards images people deems as worth sharing (e.g. well lit and framed). In contrast, person detection systems are deployed, and expected to work, in more challenging settings, such as low lighting. Due to this gap, a model that achieves high test accuracy may end up performing poorly in real world scenarios, after the design stage has ended.

To address this issue, we present a benchmark suite for TinyML person detection that tests a model's robustness in challenging settings, enabling better analysis at the design stage. Additionally the benchmark suite measures the model's performance across demographics, to ensure the model does not exhibit clear bias. The suite consists of five fine-grain benchmark sets, which we elaborate on below. Each of the five fine-grained benchmarks have been chosen based on a combination of its relevance to TinyML use cases and the availability of requisite metadata to generate the sets. Each benchmark set is a subset of the validation or test set filtered based on the criteria under test (*e.g.* persons far away from the camera).

The typical usage of the benchmark is to determine if a model is sufficiently accurate in the planned deployment setting. For example, a model designer may make different design choices for a use case where the subject is close to the camera and well lit, vs. the inverse setting. In this section, we describe the design of each individual benchmark and use the suite to compare a VWW model to a Wake Vision model. Tab. 6 reports the number of samples in each of the individual benchmark sets and the F1 scores of a Wake Vision and VWW model on each set.

4.1 Perceived Gender and Age

The Perceived Gender and Age fine grained datasets are generated from the Open Images More Inclusive Annotations for People (MIAP) extended fairness labels [32]. Underrepresented sub-groups typically constitute a small portion of a generic test set; therefore, top-line metrics often obscure a bias in a model until it is deployed [16]. This benchmark aims to evaluate a model separately on demographics that are underrepresented in the underlying dataset distribution, identifying bias. These labels are based on a perceived gender and age representation, and are not necessarily representative of a persons gender identity or age.

Table 6: Wake Vision Fine Grain Benchmark Suite. We report the samples in each set as well as the F1 score of a Wake Vision model and a VWW model on the test set.

| | | Set Size | | F1-Score | |
	Benchmark	Val	Test	Wake Vision	VWW
Gender	Female	675	2,152	**0.94**	0.89
	Male	1,238	3,748	**0.93**	0.89
	Unknown	1,718	5,231	0.78	**0.79**
Age	Young	271	875	**0.94**	0.90
	Middle	2,035	6,371	**0.93**	0.89
	Older	91	271	**0.96**	0.90
	Unknown	1,388	4,120	0.73	**0.76**
Distance	Near	5,888	17,536	**0.90**	0.88
	Medium	2,304	7,168	**0.83**	**0.83**
	Far	1,152	3,328	0.39	**0.49**
Lighting	Dark	3,036	8,872	**0.88**	0.83
	Normal	12,992	39,327	**0.88**	0.86
	Bright	936	3,083	**0.85**	**0.85**
Depictions	Person	179	501	**0.83**	0.77
	Non-Person	313	957	**0.89**	0.85
	No Depiction	7,990	24,183	**0.90**	0.87

Like the authors of the original MIAP labels, we do not support or condone the creation or deployment of gender or age specific classifiers based on this dataset.

4.2   Distance

The distance fine grained datasets aim to test how the distance of persons in images impact the performance of ML systems. This benchmark is critical for correctly predicting a model's deployed performance. If a person detection system is intended to recognize subjects at great distances, the system's performance on the far away dataset will be more informative than its performance on the top-level test set. We create three datasets based on the percentage of the image the subject bounding box covers. The three sets are near ($>$60%), at a medium distance (10-60%), and far away ($<$10%).

4.3   Lighting

The lighting fine grained datasets aim to test the performance of ML systems across different lighting conditions. For instance, in scenarios such as surveillance cameras, outdoor robotics, or augmented reality applications, models need to be robust to varying lighting conditions, including low-light environments. As with distance, this benchmark provides insight into possible causes of significant performance degradation once deployed, such as low-light settings. We create three fine grained datasets of this type for dark, normal and bright lighting conditions respectively. All images in the original Wake Vision Vision validation

and test set are classified into these datasets according to their lighting condition. We quantify lighting conditions by the average pixel values of images in greyscale, which is a simple but effective method for distinguishing lighting conditions [37]. We define low as an average pixel value less than 85, normal as between 85 and 170, and bright lighting conditions as greater than 170.

4.4   Depictions

A particularly challenging task for a person detection model is to correctly reject depictions of people. In many use cases a person detection model can not falsely trigger on a depiction. For example, a room occupancy detector that incorrectly identifies a painting on the wall as a person. This benchmark measures the models accuracy on three related sets of non-person samples: depictions of people, depictions that are not of people, and images do not contain a depiction of any kind. Depictions of people can range from photo-realistic to crude stick figures.

4.5   Benchmark Results

Tab. 6 compares the results of a model trained on the Wake Vision (Quality) training set to an identical model trained on VWW. The Wake Vision model exhibits superior robustness across the challenging settings exercised by our benchmarking suite, further demonstrating the effectiveness of the dataset. For instance, on the "Depictions" benchmark, which evaluates performance on images containing persons, non-person objects, or no depictions, the Wake Vision model achieves an F1 score of 0.83 for person depictions, outperforming the VWW model's 0.77. Similarly, for non-person depictions, the Wake Vision model scores 0.89 compared to the VWW model's 0.85.

The Wake Vision model also showcases improved performance on the "Age" benchmark, with F1 scores of 0.94 for young subjects, 0.93 for middle-aged, and a notable 0.96 for older individuals, surpassing the VWW model's scores of 0.90, 0.89, and 0.90, respectively. This highlights the Wake Vision dataset's effectiveness in enhancing model robustness for detecting people across various age groups, particularly the elderly demographic. In addition to age and depictions, the Wake Vision model demonstrates resilience to challenging lighting conditions, achieving F1 scores of 0.88 for both dark and normal lighting scenarios, compared to the VWW model's 0.83 and 0.86, respectively.

Collectively, these results underscore the significance of Wake Vision's large-scale, diverse data and fine-grained benchmarks, enabling the development of more robust and reliable TinyML person detection models across a wide range of realistic and real-world scenarios.

5   Model Design Case Study

Basic test set performance can often misrepresent a model's performance, given the typical domain shift between images scraped from the internet and real-world use cases. This issue is exacerbated when ML practitioners must trade off
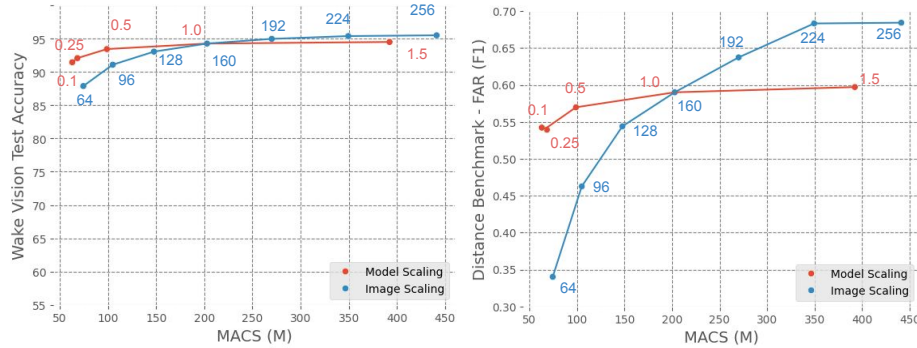
Fig. 3: Effects of scaling the image size vs. the model width on Wake Vision test accuracy and the Distance-Far benchmark set. The Distance-Far F1 score (right) is far more sensitive to changes in the image size than the overall test accuracy (left). Image sizes refer to square image resolutions (*e.g.* 96x96) and model size refers to the MobileNetV2 width multiplier. When not specified the image size is 160 and the width multiplier is 1.0.

accuracy for model performance and size, which is necessary for TinyML use cases. A design decision might have seemingly little impact on the test accuracy but may destroy real-world performance depending on the deployment environment. For example, a person detection system may only operate in dark lighting conditions, but the test dataset has an insignificant number of dark samples; therefore the test accuracy will not reflect the real world accuracy.

The benchmark suite enables more holistic analysis during the design phase. To show this use we perform a series of scaling experiments employing typical TinyML compression techniques and identifying under which circumstances these techniques are appropriate. While these results can inform ML practitioners, our intention is to demonstrate the usefulness of the benchmark suite.

5.1   Image Size vs. Model Width Scaling

We train two series of MobileNetV2 models: one series that sweeps the input image size [64-256] and one that sweeps the width multiplier of a model [0.1-1.5]. We then benchmark these models on the Wake Vision test set as well as the far distance benchmark. We plot these results against the number of multiply accumulate (MAC) operations in the model as a proxy for on-device latency [2].

The results in Fig. 3 (left) show that when looking exclusively at the high-level metric (i.e., test accuracy), scaling the input image size has a similar impact as scaling the model size. However, as shown in Fig. 3 (right), when we consider only samples where the person is far away from the camera (i.e., the distance benchmark), we observe a much more significant impact when scaling the image size. In the case of distant subjects, the image size becomes the bottleneck.

These findings suggest that for ML developers targeting use cases where the subject is likely to be far from the camera, prioritizing larger input image

sizes over wider models may be more beneficial. However, this critical design consideration could be obscured when solely relying on high-level metrics like overall test accuracy. The distance benchmark in Wake Vision effectively unveils the disproportionate impact of image size on model performance for distant subjects, enabling more informed decision-making during model optimization.

5.2   Quantization

Quantization is a crucial technique for deploying efficient TinyML models, offering substantial benefits in terms of reduced latency, memory footprint, and model size. However, prior work has suggested that quantization can disproportionately impact the performance of models on underrepresented subsets of data [17]. To assess the implications of quantization in the context of person detection, we investigate the impacts of int8 quantization on a model's benchmark results across Wake Vision's fine-grained benchmarks.

Our findings show negligible degradation in performance across all benchmarks (±0.004 F1) when employing int8 quantization, even on outlier sets. This result contradicts the previously observed disproportionate impact of quantization on underrepresented subsets. We speculate that person detection may be a relatively simple task, potentially explaining why we do not observe this specific property of quantization in our experiments. Given the negligible performance degradation and the substantial latency, memory, and model size benefits of quantization, we conclude that quantization is a win for person detection.

5.3   Grayscale

Converting a model's input image channels to grayscale from RGB is a commonly employed optimization in the TinyML field [2] as it can substantially reduce a model's memory consumption. We observed, however, that the grayscale optimization disproportionately impacts images on the brighter end of the spectrum as illustrated in Fig. 4. This further demonstrates the importance of fine grained analysis, as some real world deployment environments might be far brighter than the average Wake Vision test sample.

6   Ethical Considerations

The primary purpose of the Wake Vision dataset and benchmark suite is to mature the field of TinyML by giving researchers access to large-scale datasets and enabling fine-grain analysis. Beyond this purpose, Wake Vision can directly improve the quality of person detection systems that benefit society, such as saving energy by only turning on the lights in a room when a person is present. Additionally, TinyML person detection protects user privacy by running inference on-device and not transmitting private data over the network. However, it should be noted that person detection models can also be used for malicious purposes, such as weapons targeting systems or large-scale surveillance.
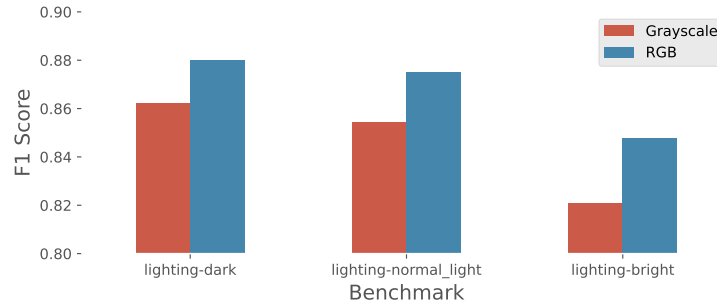
Fig. 4: Impact of grayscale input images on the lighting benchmarks: dark, normal light, and bright. Models that use grayscale input images are more sensitive to bright lighting conditions than RGB.

The images that make up Wake Vision are entirely sourced from Flickr. While we attempt to only use images under the consent given by the CC-BY 2.0 license (by sourcing images through Open Images [19, 21]), we make no representations or warranties regarding the license status of each image. It is possible that some of the millions of images have been uploaded to Flickr by a user without the right to distribute the images with a CC-BY 2.0 license.

Futhermore, the benchmark suite is intended to ensure the fairness and robustness of person detection models. While the goal of these benchmark suites is to build more robust and fair models, our demographic benchmarks could be misused to create systems to classify gender and age, potentially breaching personal privacy or perpetuating discrimination. We have attempted to mitigate this risk by not releasing training sets for the fine-grained benchmark datasets.

7   Conclusions

The field of TinyML research is currently bottlenecked by the lack of large, high-quality datasets. To alleviate this bottleneck we introduced Wake Vision—a large scale dataset and benchmark suite for person detection. Compared to VWW, the current state of the art dataset for person detection, Wake Vision is ~100x larger, easier to use, has a comparable amount of label errors, and is more reflective of real world use cases. These factors contribute to Wake Vision being a drop-in improvement over VWW. Fairness across demographics and robustness to challenging settings is essential for many person detection use cases. To facilitate evaluation of this, we released a Wake Vision benchmark suite of six fine-grained benchmark sets to test the fairness and robustness of person detection models. Lastly we demonstrate how the Wake Vision benchmark suite can be used to investigate the effects of different types of input and model scaling and highlight the importance of fine grain analysis when designing an model.

References

1. Banbury, C., Reddi, V.J., Torelli, P., Holleman, J., Jeffries, N., Kiraly, C., Montino, P., Kanter, D., Ahmed, S., Pau, D., et al.: Mlperf tiny benchmark. arXiv preprint arXiv:2106.07597 (2021) 1, 2, 3

2. Banbury, C., Zhou, C., Fedorov, I., Matas, R., Thakker, U., Gope, D., Janapa Reddi, V., Mattina, M., Whatmough, P.: Micronets: Neural network architectures for deploying tinyml applications on commodity microcontrollers. Proceedings of Machine Learning and Systems **3**, 517–532 (2021) 2, 3, 12, 13

3. Banbury, C.R., Reddi, V.J., Lam, M., Fu, W., Fazel, A., Holleman, J., Huang, X., Hurtado, R., Kanter, D., Lokhmotov, A., et al.: Benchmarking tinyml systems: Challenges and direction. arXiv preprint arXiv:2003.04821 (2020) 1

4. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020) 6

5. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10925–10934 (2022) 6

6. Blalock, D., Gonzalez Ortiz, J.J., Frankle, J., Guttag, J.: What is the state of neural network pruning? Proceedings of machine learning and systems **2**, 129–146 (2020) 3

7. Chowdhery, A., Warden, P., Shlens, J., Howard, A., Rhodes, R.: Visual wake words dataset. arXiv preprint arXiv:1906.05721 (2019) 2, 3, 8, 21

8. David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Wang, T., et al.: Tensorflow lite micro: Embedded machine learning for tinyml systems. Proceedings of Machine Learning and Systems **3**, 800–811 (2021) 1, 3

9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 2

10. Evci, U., Gale, T., Menick, J., Castro, P.S., Elsen, E.: Rigging the lottery: Making all tickets winners. In: International Conference on Machine Learning. pp. 2943–2952. PMLR (2020) 3

11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html 3

12. Flickr: https://flickr.com/ (2024), accessed: 2024-03-06 4

13. Gale, T., Elsen, E., Hooker, S.: The state of sparsity in deep neural networks. arXiv preprint arXiv:1902.09574 (2019) 3

14. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded classification models: Combining models for holistic scene understanding. Advances in neural information processing systems **21** (2008) 2

15. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 2

16. Hooker, S., Courville, A., Clark, G., Dauphin, Y., Frome, A.: What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248 (2019) 3, 9

17. Hooker, S., Moorosi, N., Clark, G., Bengio, S., Denton, E.: Characterising bias in compressed models. arXiv preprint arXiv:2010.03058 (2020) 3, 13

18. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2704–2713 (2018) 3

19. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017) 2, 14

20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Masters Thesis (2009) 3

21. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020) 2, 14

22. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.: Datasets: A community library for natural language processing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-demo.21 8

23. Lin, J., Chen, W.M., Lin, Y., Gan, C., Han, S., et al.: Mcunet: Tiny deep learning on iot devices. Advances in Neural Information Processing Systems 33, 11711–11722 (2020) 3

24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 3

25. Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H.R., et al.: Dataperf: Benchmarks for data-centric ai development. Advances in Neural Information Processing Systems 36 (2024) 6

26. Njor, E., Madsen, J., Fafoutis, X.: Data aware neural architecture search. arXiv preprint arXiv:2304.01821 (2023) 2, 6

27. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research 70, 1373–1411 (2021) 7

28. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749 (2021) 6, 7

29. Piechocki, M., Kraft, M., Pajchrowski, T., Aszkowski, P., Pieczynski, D.: Efficient people counting in thermal images: The benchmark of resource-constrained hardware. IEEE Access 10, 124835–124847 (2022) 2

30. Riccardi, G., Hakkani-Tur, D.: Active learning: Theory and applications to automatic speech recognition. IEEE transactions on speech and audio processing 13(4), 504–511 (2005) 6

31. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: In-verted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018) 5, 22

32. Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., Pantofaru, C.R.: A step toward more inclusive people annotations for fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (2021) 9

33. TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets 8

34. Warden, P.: Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018) 2

35. Warden, P., Situnayake, D.: Tinyml: Machine learning with tensorflow lite on ar-duino and ultra-low-power microcontrollers. O'Reilly Media (2019) 1

36. Zacharia, A., Zacharia, D., Karras, A., Karras, C., Giannoukou, I., Giotopoulos, K.C., Sioutas, S.: An intelligent microprocessor integrating tinyml in smart hotels for rapid accident prevention. In: 2022 7th South-East Europe Design Automa-tion, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). pp. 1–7. IEEE (2022) 2

37. Zhang, W., Li, H., Wang, Z.: Research on different illumination image classifica-tion method. In: 2017 2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017). pp. 574–581. Atlantis Press (2017) 11

A   Anonymized Repo

The code used for the paper is available at this anonymous repository: https://anonymous.4open.science/r/Wake_Vision-0D15/README.md

This repo contains the code to generate Wake Vision and the benchmark suite, as well as the code to train and evaluate models. This code is sufficient to reproduce all results in the paper.

B   Flowchart of Bounding Box Filtering Process

Fig. 5 illustrates the filtering process for Wake Vision when using Open Images' bounding box labels as the label source.

C   Fine Grained Benchmark Images

Fig. 6 gives example images for each of the benchmarks in the suite.

D   Open Image Label Distribution

Wake Vision is derived from Open Images V7, therefore the diversity of subjects in the images should follow the distribution of labels in Open Images. The Label distribution of Open images can be found here: https://storage.googleapis.com/openimages/web/factsfigures_v7.html#statistics

E   Dataset Access and Organization

Wake Vision is available directly online, via TFDS, or Hugging Face Datasets. In order to maintain author anonymity, we do not link to the dataset.

The dataset is organized as a set of compressed tar files containing images, and a series of label CSVs. The label CSVs are organized such that the file name of the image is the identifying index. The person label is provided as is a flag to indicate if the image is of a depiction. The Validation and Test label CSVs also have flags that denote a sample's inclusion into a fine-grain benchmark set (e.g. Distance-Near). This structure makes it easy to access just the required data without requiring a full download. It also ensures the dataset can be easily updated as new versions are introduced.

F   Label Generation Details

Person Labels. The most straightforward Open Images label classes to label as person in Wake Vision are the "person" label and its subcategories (listed in Appendix I). All of these are relabelled as persons in Wake Vision. These labels are present as both image-level labels and bounding box labels.
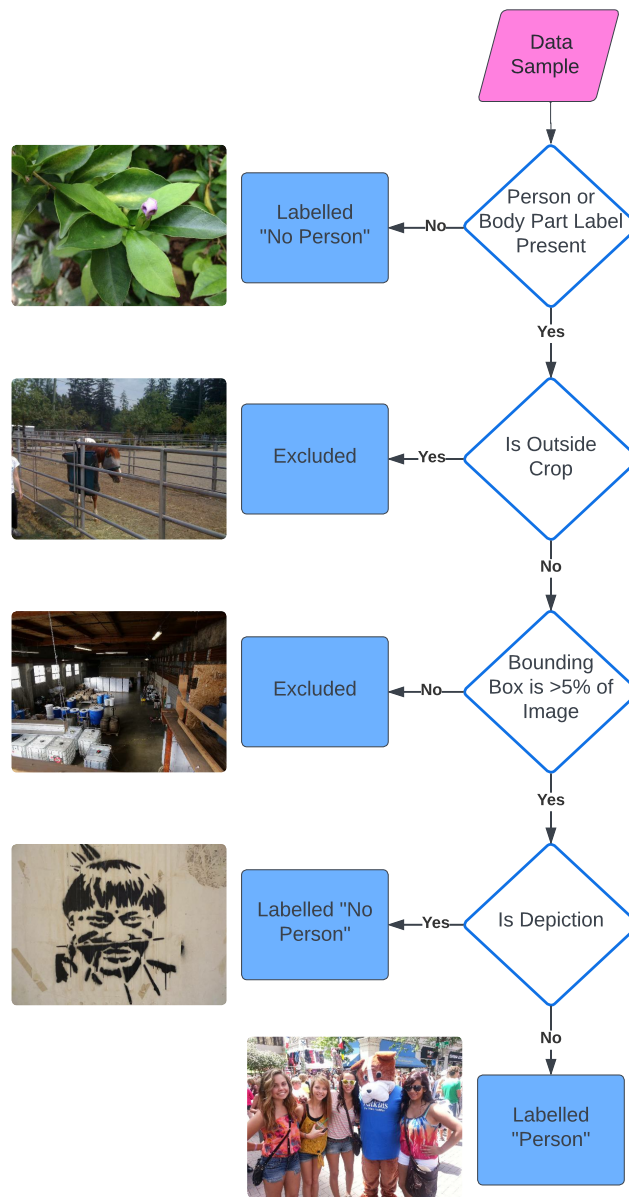
Fig. 5: A flowchart of the filtering of an image according to default Bounding Box filtering rules.
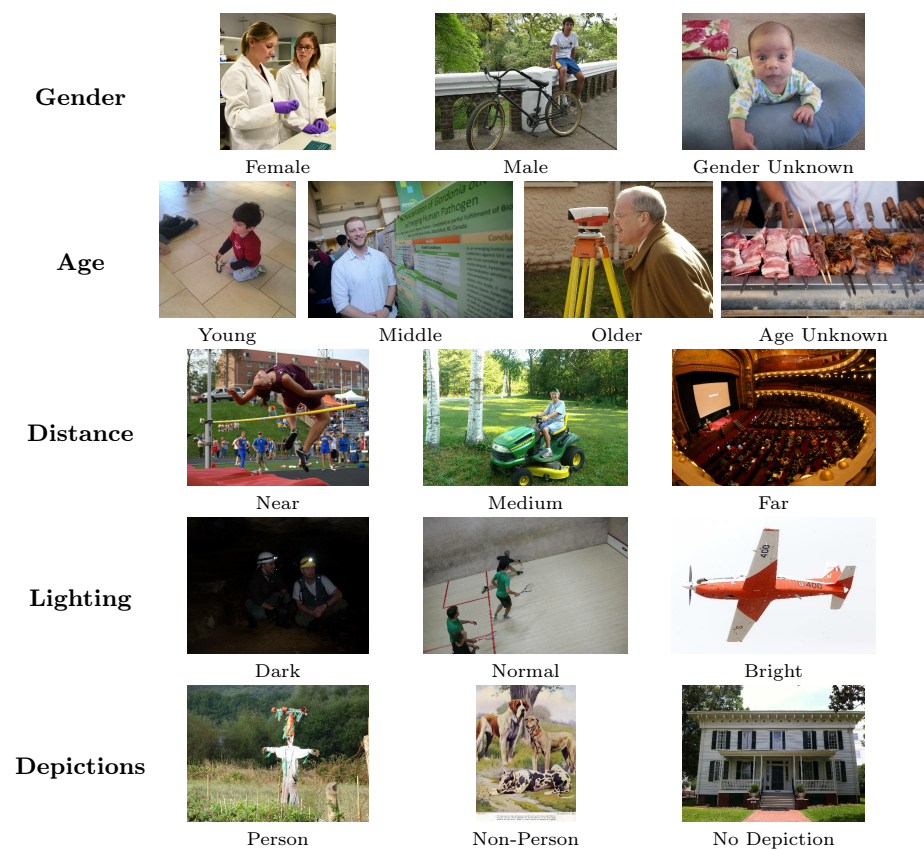
Fig. 6: Images from each fine grained benchmark dataset.

We furthermore inspect the image level label classes for synonyms and umbrella terms for all the person related labels. This search resulted in an additional six person related label classes. These person related label classes will only be used in the image level label configuration.

Person Body Part Labels. Body parts are more challenging to relabel, as it is dependent on the use case whether a body part should be considered a person.

For example a camera that detects whether a person is inside a room to decide if the light should be switched on would want to consider body parts as a person, as this will keep the lights on even when the person is only partly in the camera frame. For waking up electronics, however, it may not make sense to consider body parts as persons. This could, *e.g.*, mean that a computer would turn on when detecting a foot.

To cater to both use cases we include a flag in our open-source dataset creation code to set whether body parts should be considered persons. By default we consider body parts as persons.

Depictions. Open Images bounding box labels contain metadata about whether an object is a depiction, *e.g.*, a painting or a photograph of a person. This presents the challenge of how to handle depictions. While most use cases would not consider a depiction a person, it could make sense to either exclude them to make training easier, or include them as non-persons to make a model resistant to seeing depictions when deployed.

In line with how we handle body parts, we therefore include a flag for our open-source dataset creation code to set whether depictions should be excluded or considered non-persons. By default depictions are considered non-persons.

Bounding Box Size. The VWW dataset only considered a Common Objects in Context (COCO) person to be a person if the bounding box around the person took up at least 5% of the image [7]. If the person took up less than 5% of the image, the image was excluded from the dataset. To make our dataset work as a plug in replacement for VWW, we adopt the same defaults in Wake Vision. For different requirements, users can change a configuration parameter in our open-source dataset creation code.

G   Standardized Evaluation

VWW has no standardized way to evaluate performance on the dataset. This makes it challenging to compare works based on the dataset, since performance difference can come down to choices outside the contribution of the work. *E.g.*, two works could be contributing with model improvements, but use different pre-processing pipelines that skews results.

To allow for both data-centric and model-centric improvements, we provide a standard model for data-centric contributions, and a standard pre-processing pipeline for model-centric contributions. Both types of contributions are expected to use accuracy as the primary metric for the overall test and valida-

tion set, and F1-score as the primary metric for the fine grained benchmarks introduced in Sec. 4.

Therefore, for data-centric contributions we propose to use a Mobilenet v2 model with a width modifier* of 0.25 [31]. For model-centric contributions we propose the following pre-processing pipeline:

1. Cast image pixel value datatype from 8-bit integers into 32-bit floating points
2. Resize image such that the shortest side matches the model input size
3. Perform a center crop on the image such that the longest size matches the model input size
4. Normalize pixel values to between -1 and 1 sample wise
5. Use image tensor as input features and person label as target feature

H   Dataset Download

The full Open Images v7 dataset is not hosted by the dataset authors. Rather it is provided as a collection of flickr image Uniform Resource Locators (URLs) and their associated labels. As an unfortunate result of this, the dataset is not static over time as image owners can delete their images from the flickr platform. As a result we were only able to download a subset of the original Open Images v7 dataset as shown in Tab. 7.

Table 7: Number of images downloaded from Open Images v7. Download occurred between the 28th of November to the 5th of December

	Train	Validation	Test
Downloaded	7,936,979	36,406	109,305
Errors	1,055,669	5,214	16,131

I   Person Label Classes

We consider the following Open Images v7 labels to be a person for the Wake Vision dataset:

- *Person*
- *Woman* (Subcategory of *Person*)
- *Man* (Subcategory of *Person*)
- *Girl* (Subcategory of *Person*)
- *Boy* (Subcategory of *Person*)
- *Human body* (Part of *Person*)

---

* Also known as alpha.

- *Human face* (Part of *Person*)
- *Human head* (Part of *Person*)
- *Human* (Person synonym - Only in Image Level Label Configuration)
- *Female person* (Woman synonym - Only in Image Level Label Configuration)
- *Male person* (Man synonym - Only in Image Level Label Configuration)
- *Child* (Umbrella term for *Girl* & *Boy* - Only in Image Level Label Configuration)
- *Adolescent* (Umbrella term for *Girl* & *Boy* - Only in Image Level Label Configuration)
- *Youth* (Umbrella term for *Girl* & *Boy* - Only in Image Level Label Configuration)

Images containing the following Open Images v7 labels and no other person related labels are excluded from the Wake Vision dataset:

- *Human eye* (Part of *Person*)
- *Skull* (Part of *Person*)
- *Human mouth* (Part of *Person*)
- *Human ear* (Part of *Person*)
- *Human nose* (Part of *Person*)
- *Human hair* (Part of *Person*)
- *Human hand* (Part of *Person*)
- *Human foot* (Part of *Person*)
- *Human arm* (Part of *Person*)
- *Human leg* (Part of *Person*)
- *Beard* (Part of *Person*)

J   Wake Vision Dataset Size

Table 8: Amount of images in the Wake Vision dataset

	Person Images	Non-Person Images	Excluded
Image Level Label Train Dataset	2,880,214	2,880,214	2,176,551
Bounding Box Label Train Dataset	624,115	624,115	6,688,749
Validation Dataset	8,482	8,482	19,442
Test Dataset	25,641	25,641	58,023