This ACCV 2022 workshop paper, provided here by the Computer Vision Foundation, is the author-created version. The content of this paper is identical to the content of the officially published ACCV 2022 LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

CaltechFN: Distorted and Partially Occluded Digits

Patrick Rim[1], Snigdha Saha[1], and Marcus Rim[2]

[1] California Institute of Technology, Pasadena, CA 91125, USA
{patrick,snigdha}@caltech.edu
[2] Vanderbilt University, Nashville, TN 37235, USA
marcus.g.rim@vanderbilt.edu

Abstract. Digit datasets are widely used as compact, generalizable benchmarks for novel computer vision models. However, modern deep learning architectures have surpassed the human performance benchmarks on existing digit datasets, given that these datasets contain digits that have limited variability. In this paper, we introduce Caltech Football Numbers (CaltechFN), an image dataset of highly variable American football digits that aims to serve as a more difficult state-of-the-art benchmark for classification and detection tasks. Currently, CaltechFN contains 61,728 images with 264,572 labeled digits. Given the many different ways that digits on American football jerseys can be distorted and partially occluded in a live-action capture, we find that in comparison to humans, current computer vision models struggle to classify and detect the digits in our dataset. By comparing the performance of the latest task-specific models on CaltechFN and on an existing digit dataset, we show that our dataset indeed presents a far more difficult set of digits and that models trained on it still demonstrate high cross-dataset generalization. We also provide human performance benchmarks for our dataset to demonstrate the current gap between the abilities of humans and computers in the tasks of classifying and detecting the digits in our dataset. Finally, we describe two real-world applications that can be advanced using our dataset. CaltechFN is publicly available at `https://data.caltech.edu/records/33qmq-a2n15`, and all benchmark code is available at `https://github.com/patrickqrim/CaltechFN`.

1 Introduction

The task of classifying digits was one of the first computer vision tasks successfully “solved” by deep learning architectures. Released in 1998, the MNIST dataset [1] serves as a benchmark for model performance in the task of classifying digits. However, deep learning models have been able to achieve human levels of performance in the task of classifying the digits in the MNIST dataset [2–5]. Due to the standardized nature of the handwritten digits in MNIST, there is low variability between the digits, which makes it easy for modern computer vision architectures to learn the characteristic features of each digit [6, 7].

The Street View House Numbers (SVHN) dataset [8] consists of digits from house numbers obtained from Google Street View images, which pose a more difficult challenge than MNIST. Due to the natural settings and diversity in the designs of the house numbers, there is a far higher variability between the digits in SVHN than in MNIST. Thus, when SVHN was published in 2011, there was initially a large gap between human performance and model performance in the task of classifying its digits [9]. Because of this disparity, SVHN began to serve as a more difficult benchmark for novel image classification and object detection models [10,11]. However, newer models have since been able to achieve a classification accuracy on SVHN exceeding 98% [12–15], which is the published human performance benchmark. Some recent models have even achieved an accuracy exceeding 99% [16,17]. With minimal room left for improvement in performance on SVHN, there is a need for a more difficult digit dataset to benchmark the progress of future classification and detection models.

In this paper, we present **Caltech F**ootball **N**umbers (CaltechFN), a new benchmark dataset of digits from American football jerseys. Samples of the digits in our dataset are displayed in Fig. 1. We demonstrate that the latest image classification and object detection models are not able to achieve human performance on our dataset. This performance gap can be explained by the significantly increased variability of CaltechFN compared to current benchmark digit datasets. Due to the nature of American football jerseys, many of the digits in the dataset are wrinkled, stretched, twisted, blurred, unevenly illuminated, or otherwise distorted [18–21]. Sample images containing distorted digits are displayed in Fig. 2(a). These possibilities introduce a substantial number of ways that each digit can differ in appearance from the other digits in its class. We demonstrate that even the latest models struggle to learn the characteristic features of each digit when trained on such highly variable digits. This is likely due to the scarcity of digits that are distorted and occluded in the same way. As improved few-shot learning methods are developed, we expect an improvement in model performance on our dataset.

Furthermore, due to the nature of American football games, many images of digits on jerseys in live-action will be partially occluded [18–20]. For example, another player or the ball may be present between the camera and the subject player, or the player may be partially turned away from the camera such that parts of digits are not visible. Sample images containing partially occluded digits are displayed in Fig. 2(b). CaltechFN contains many such images of digits that are partially occluded, yet identifiable by human beings. This can be explained by recent neuroscience studies that have demonstrated the capability of the human brain to "fill in" visual gaps [22–25]. On the other hand, computer vision models struggle to fill in these visual gaps [26, 27] since they are often unique and not represented in the training set. In other words, there are a large number of unique ways in which a certain digit may be partially occluded. Compounding this with the number of ways that a digit can be distorted, it is difficult for models trained on our dataset to learn the characteristic features of each digit [28].

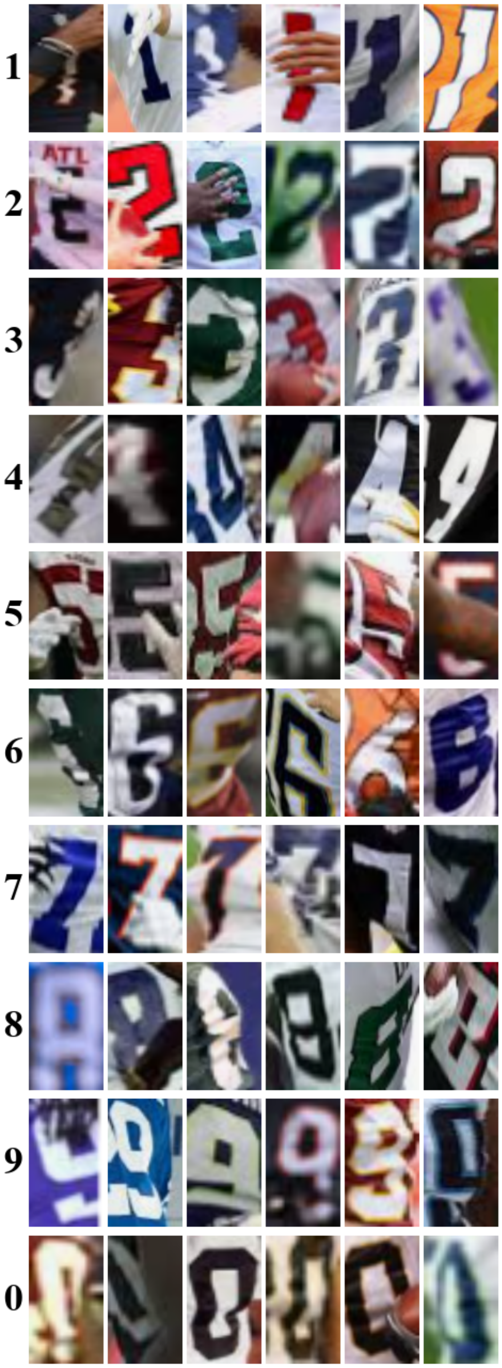

Fig. 1: Samples of cropped digits from the CaltechFN dataset that are distorted (e.g. wrinkled and stretched) and partially occluded.

The main contributions of this paper are as follows:

1. We present CaltechFN, a dataset of distorted and partially occluded digits. The CaltechFN dataset poses a difficult challenge for even the latest computer vision models due to its high intra-class variability. For this reason, CaltechFN can serve as a state-of-the-art benchmark for future image classification, object detection, and weakly supervised object detection (WSOD) models.
2. We perform experiments to measure cross-dataset model performance benchmarks on the CaltechFN and SVHN datasets. The results illustrate that CaltechFN is indeed a more difficult benchmark than SVHN and that models trained on CaltechFN demonstrate high cross-dataset generalization.
3. We record human performance benchmarks on the CaltechFN dataset using experiments with human volunteers. The existing gap between the best current model performance and our human performance benchmark will hopefully catalyze innovations in the construction of computer vision models.

This paper is structured as follows: Section 2 compares and contrasts CaltechFN with related datasets. The properties and goals of our dataset are introduced in Section 3. This section also describes the process undertaken to construct the dataset. Section 4 details and compares the performance of various image classification, object detection, and WSOD models on our dataset. We then provide human performance benchmarks on our dataset in Section 5. In Section 6, we present examples of real-world tasks that can be better solved by models trained on our dataset. In Section 7, we discuss future directions that can be taken to utilize the richness of information in the images in our dataset.

2 Related Work

Digit Datasets. Digit datasets are advantageous in their simplicity and their ease of use, containing a small number of classes and requiring little preprocessing and formatting to begin the training process. While ImageNet [29], MS-COCO [30], and PascalVOC [31] are the most popular datasets for image classification and object detection, they lack the compactness and the simplicity of digit datasets due to their large size and wide variety of classes. There are several popular digit datasets available, but many of the digits in these datasets are handwritten. MNIST [1] was the first prominent digit dataset, but the standardization of the digits limits variability. The ARDIS dataset [32] contains handwritten digits from old Swedish church records, which introduces some variability due to age-induced weathering. However, the variability is still limited by the standardized nature of handwritten digits. There do exist datasets that instead contain digits in natural settings. Roughly 10% of the Chars74k dataset [33] is from real-life, outdoor images. However, Chars74k also contains non-digit characters, which limits the number of digits it contains. SVHN [8] is the primary dataset consisting exclusively of digits in real-world settings. However, the images in SVHN have few distortions besides natural blur, and close to no occlusions since they are house numbers intended to be visible from the street. Meanwhile, CaltechFN consists of many examples of distorted and partially occluded digits, which constitutes a more difficult set of digits than any existing digit dataset.

Datasets with Distorted Objects. Distortions in many popular image datasets are limited in their complexity. For instance, the SmartDoc-QA dataset [34] contains images of documents distorted by blur, perspective, and illumination effects. All of these distortions fall under the same general domain and do not cover the wide variety of distortions in the real world. The dataset of soccer jersey numbers by Gerke, Müller, and Schäfer [35] consists of images taken from soccer videos with image-level annotations of jersey numbers. The distortions in these images are similar to the ones found in CaltechFN since they were also captured from sports settings. However, unlike CaltechFN, this dataset does not contain bounding box annotations, meaning that the dataset cannot be used to benchmark object detection tasks. Furthermore, there is more physical contact between players in American football than in soccer, meaning that there are more distorted digits in CaltechFN than in the soccer dataset, which can be empirically confirmed when observing the two datasets.

Datasets with Partially Occluded Objects. There are also many existing datasets with partially occluded objects. However, most of these datasets are not focused on digits, but on a larger range of objects. For example, the Occluded REID [36] and the Caltech Occluded Faces in the Wild [37] datasets present the challenge of identifying humans and faces, respectively, when partially occluded by other objects. Similarly, the Pascal3D+ dataset [38] augments images from the PascalVOC and ImageNet datasets with 3D annotations, partially occluding the target objects. All of these datasets lack the simplicity and convenience of digit datasets. Chars74k does contain some digits and characters that are partially occluded. However, unlike CaltechFN, the Chars74k dataset does not provide bounding boxes and thus cannot be used to benchmark object detection tasks.

3 Caltech Football Numbers (CaltechFN) Dataset

3.1 Dataset Construction

Image Collection. The first step of the data construction process was to collect candidate images. In order to construct a representative and unbiased dataset, we chose to sample an equal number of images of each jersey number in American football, which ranges from 1 to 99. Since each digit from 0 to 9 is roughly equally represented in this range, we sampled uniformly across each jersey number. This is to ensure that models trained on our dataset do not overfit to any over-represented digits [39].

We collected our candidate images by querying the Google Image Search database. Using the query "Team Name" + "Number", we collected 50 images

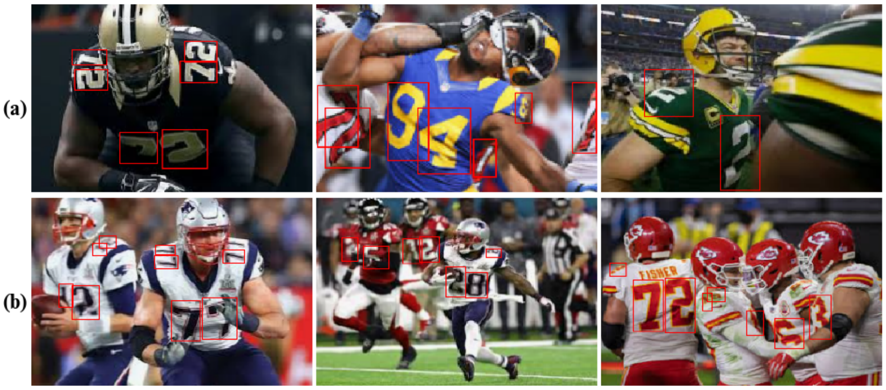

Fig. 2: Samples of full images from the CaltechFN dataset that contain labeled bounding boxes around digits that are **(a)** distorted (e.g. wrinkled and stretched), and **(b)** partially occluded.

for each combination of the 32 teams in the National Football League and the 99 jersey numbers, for a total of 1600 images for each number and a grand total of 158,400 images. We chose this query because it seemed to be neither too general nor too specific: queries that were too general returned too many irrelevant images, while queries that were too specific did not return a sufficient number of relevant images.

Image Filtering. We then completed a filtering process to remove unwanted and duplicate images. First, we made two full passes through the set of candidate images to remove images that did not contain any digits. Almost half of the candidate images were removed in this step. Then, we removed any images where the digits contained were distorted or occluded to the extent that we were not able to identify them. We first made another two full passes through the set of candidate images to identify and mark images that contained digits that were not immediately identifiable. We then carefully sorted through each of these marked images, only keeping those that contained at least one identifiable digit. Finally, we used a deduplication tool [40] to identify and remove duplicate images. The result of this cleaning process is our current set of 61,728 images and 264,572 labeled digits.

Image Annotation. As done by many previous studies [41], we utilized the Amazon Mechanical Turk (AMT) platform [42] to label each individual digit in each of the images. For each digit, including those that were partially occluded, partially cut off, or rotated, AMT workers were asked to draw and label a maximally tight bounding box that contained every visible pixel of the digit. We then worked through the results and fixed any errors; specifically, we labeled identifiable digits that were not already labeled, removed erroneous boxes, and corrected any incorrect labels.

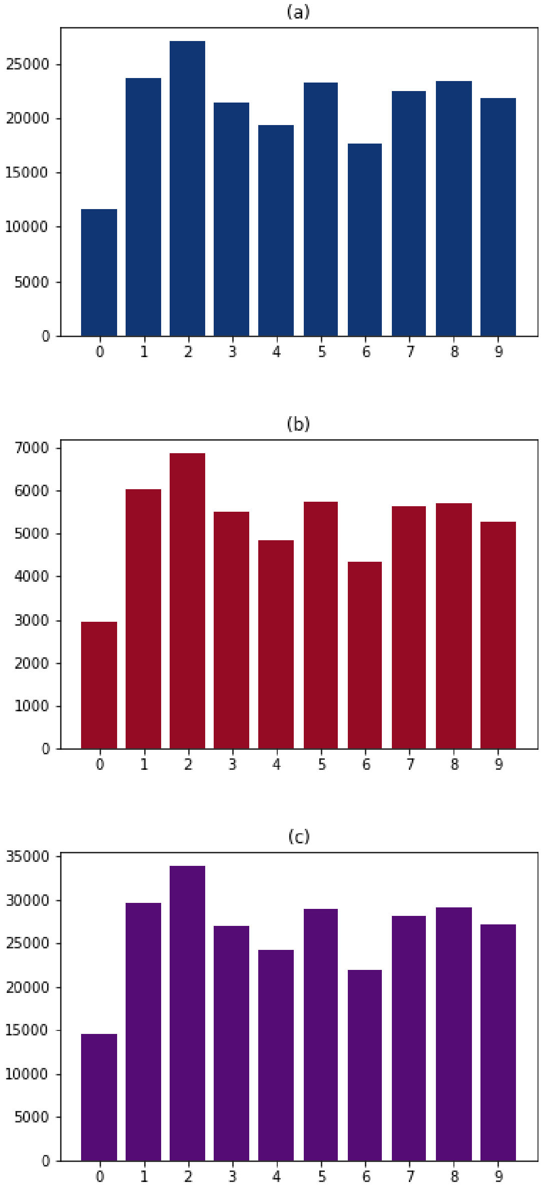

Fig. 3: Distribution of digits in **(a)** the train set, **(b)** the test set, and **(c)** the entire dataset.

3.2 Properties

Through CaltechFN, we aim to provide an extensive set of digits with high variability to provide a new goal for Computer Vision models to work towards. To that end, the dataset includes a highly variable set of digits, with many of them being distorted and partially occluded in unique ways. For instance, some digits are partially blocked by a football, while other digits are twisted and wrinkled due to the jersey being pulled on. We also include some easier images, such as stationary images of unobstructed jerseys. We hope that the variability of digits presented in the dataset will challenge researchers to design innovations that will allow models to identify similarities between digits of the same class.

We will now provide more specific details about the CaltechFN dataset. As described in Section 3.1.2, the current version of our dataset contains a total of 61,728 images and 264,572 labeled digits. As shown in Fig. 3, our dataset contains a roughly uniform number of images from each of the ten digit classes. As mentioned in Section 3.1.1, this is necessary to ensure that models trained on our dataset do not overfit to any over-represented digits.

We publish our dataset in the "Full Image" and "Cropped Digits" formats:

- The "Full Image" format includes all images in their original resolutions as obtained from the image collection process. Each image is accompanied by bounding box annotations for each identifiable digit that it contains. The mean and standard deviation of the heights and widths of the full images are 181.571±15.452 pixels and 234.692±54.365 pixels respectively. Samples of full images with the bounding boxes drawn are displayed in Fig. 2.
- The "Cropped Digits" format contains character-level images of each digit. These images were created by cropping and labeling each region of the full images contained by a bounding box. The mean and standard deviation of the heights and widths of the cropped digits are 32.360±18.042 pixels and 21.334±9.375 pixels respectively. Samples of cropped digits are displayed in Fig. 1.

For both formats of our dataset, we provide a train set ("CaltechFN-train") and a test set ("CaltechFN-test"). This train-test split was created using a random, uniform 80-20 split. As seen in Fig. 3, the distribution of digits in the train set and test set are similar to the overall distribution of digits across the entire dataset. The details of the train-test split for both the "Full Image" and "Cropped Digits" formats are as follows:

- "Full Image": train set contains 49,383 images (80.0% of total), test set contains 12,345 images (20.0% of total).
- "Cropped Digits": train set contains 211,611 digits (80.0% of total), test set contains 52,911 digits (20.0% of total).

Table 1: Model performance when **(A)** trained on CaltechFN-train, tested on CaltechFN-test; (B) trained on CaltechFN-train, tested on SVHN-test; (C) trained on SVHN-train, tested on CaltechFN-test; (D) trained on SVHN-train, tested on SVHN-test. Image classification models are evaluated using classification accuracy. Object detection and WSOD models are evaluated using mAP.

Image Classification

Model	(A)	(B)	(C)	(D)
MobileNet (CVPR '18) [43]	86.0±0.6	93.1±0.5	76.0±0.6	98.2±0.5
DenseNet121 (CVPR '17) [15]	87.9±0.4	95.0±0.3	77.9±0.3	98.6 ±0.4
ResNet50 (CVPR '16) [44]	86.9±0.4	94.2±0.6	77.1±0.5	98.3±0.4

Object Detection

Model	(A)	(B)	(C)	(D)
YOLOv5 ('21) [45]	54.4±0.5	61.2±0.5	37.5±0.9	67.9±0.4
RetinaNet (ICCV '17) [46]	52.7±0.8	57.8±0.7	30.0±1.4	65.2±0.7
SSD (ECCV '16) [47]	54.6±0.4	61.1±0.2	38.6±1.0	67.2±0.4
Faster-RCNN (NIPS '15) [48]	57.4±0.3	60.9±0.5	38.8±0.6	68.5±0.3

Weakly Supervised Object Detection (WSOD)

Model	(A)	(B)	(C)	(D)
Wetectron (CVPR '20) [49]	29.5±0.6	37.5±0.5	20.7±1.8	42.6±0.3
C-MIL (CVPR '19) [50]	26.3±1.4	36.0±1.2	17.2±1.0	39.4±0.8
WSOD2 (ICCV '19) [51]	21.1±0.4	27.0±0.8	14.8±1.8	30.9±0.3
PCL (CVPR '17) [52]	27.1±0.9	34.5±0.6	16.9±1.0	37.3±1.1

4 Model Performance

In the following experiments, we will compare the performance on CaltechFN and SVHN of some of the latest models built for the tasks of image classification, object detection, and weakly supervised object detection. We compare performance on our dataset to performance on SVHN because it is the most similar existing digit dataset, as we explained in Section 2. We will show that CaltechFN is a significantly more difficult dataset than SVHN, while also showing that models trained on CaltechFN perform at least as well as models trained on SVHN. For each model, we will present results for the following four experiments, labeled as follows:

(A) Training on CaltechFN-train, testing on CaltechFN-test,
(B) Training on CaltechFN-train, testing on SVHN-test,
(C) Training on SVHN-train, testing on CaltechFN-test,
(D) Training on SVHN-train, testing on SVHN-test.

The experimental results are presented in Table 1. We note that the results labeled (A) serve as benchmark performance scores for CaltechFN in the three tasks we detailed. The evaluation metric for the image classification results is classification accuracy, while the evaluation metric for the object detection and WSOD results is mAP. All experimental details including the hyperparameter search effort, the compute resources used, and a description of the evaluation metrics are discussed at length in the Supplementary Material.

We now explain the relevance of the experimental results:

1. **We demonstrate the comparative difficulty of CaltechFN.** When trained on CaltechFN, models perform worse on CaltechFN (A) than on SVHN (B). The same models trained on SVHN also perform worse on CaltechFN (C) than on SVHN (D). Regardless of which dataset is used as the training set, models perform worse on CaltechFN than on SVHN.

2. **We demonstrate that models trained on CaltechFN demonstrate high cross-dataset generalization.** Models perform worse on CaltechFN when training on SVHN (C) than when training on CaltechFN (A) itself, but the performance of the same models on SVHN does not significantly drop when trained on CaltechFN (B) instead of on SVHN (D). This shows that computer vision models are able to learn robust, generalizable features by training on CaltechFN.

5 Human Performance Benchmark

To demonstrate the potential for improvement in current computer vision architectures, we provide human performance benchmarks on our dataset in the same classification and detection tasks performed by computer vision models in Section 3.

To measure human performance in the task of classifying the cropped digits in CaltechFN, we asked five human volunteers to label a subset of 15,000 cropped digits ("All Samples"). We calculate mean human performance by computing the accuracies of the volunteer-generated labels.

To measure human performance in the task of detecting digits in the full images in CaltechFN, we asked the same five human volunteers to draw bounding boxes on a subset of 5,000 images ("All Samples"). We calculate mean human performance using the same mAP metric used for the object detection models in Section 3.

Furthermore, we calculate the human performance in both tasks for only the subset of "Difficult Samples" that even the best models in Section 4 were unable to classify/detect. The results, which we provide as benchmarks for human performance in the tasks of image classification and object detection on our dataset, are presented in Table 2.

We see that humans are able to achieve high levels of performance, even on the samples that the best models were unable to classify/detect. This clear disparity between human performance and the best model performances in both

Table 2: Human performance on the CaltechFN dataset. Image classification performance is evaluated using classification accuracy, while object detection performance is evaluated using mAP.

Human Performance Benchmarks

Task	All Samples	Difficult Samples
Image Classification	99.1±0.8	97.8±2.3
Object Detection	87.2±4.5	83.0±5.9

tasks demonstrates that there potentially exist certain techniques not yet learned by models that humans use to identify difficult digits. Evidently, there is still a need for innovations in the construction of computer vision models for computers to be able to achieve human levels of performance in the aforementioned tasks.

However, it is clear that even humans find it difficult to identify every digit in our dataset. Even though the digits that are included in our dataset are the ones that we approved as identifiable, it is not necessarily true that other humans will also be able to identify each of these digits. This may be due to a bias stemming from the fact that we collected the images, or simply due to variability in performance across different humans. Ultimately, this leaves open the possibility that significant advancements in computer vision techniques may result in models being able to achieve even higher performance on classification and detection tasks on our dataset than humans.

Fig. 4 provides a visual representation of Tables 1 and 2. The comparative difficulty of CaltechFN and the high cross-dataset generalization of models trained on CaltechFN, as well as the gap between human performance and model performance on CaltechFN, are clearly illustrated.

6 Applications of CaltechFN

We will now present two potential real-world applications that involve detecting and classifying digits. Our dataset provides a rich source of difficult digits that are distorted and partially occluded in the same manner that they are in these two real-world settings. We will explain the benefits of using the images in our dataset to train models that can perform these real-world tasks.

6.1 Player Detection and Tracking in Sports

Currently, coaches in sports must watch footage of a past game to chart the personnel (players on the field) over the duration of the game, which is an important task in sports analytics [19]. Computer vision models have thus far struggled to outperform humans at this task due to the distortions and occlusions of jersey numbers, which are the primary identifying features of players [18]. Our dataset can be used to train a model that can identify players in game footage using

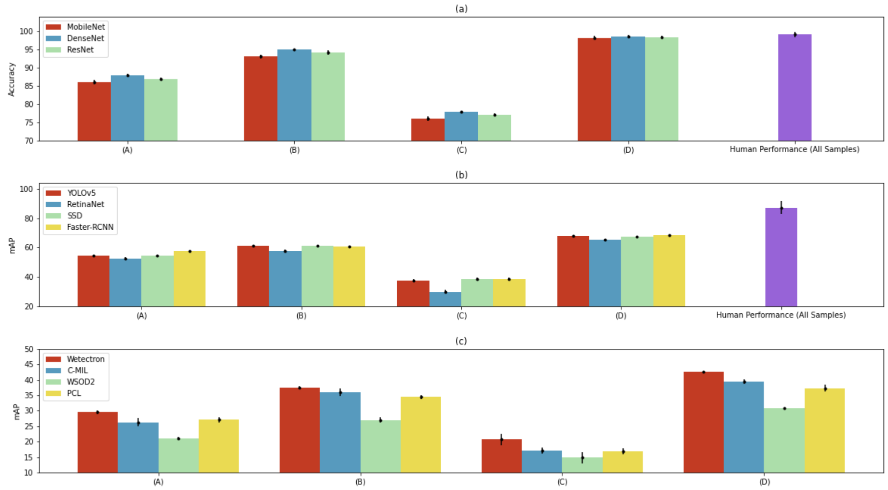

Fig. 4: Visual representation of Tables 1 and 2 for **(a)** image classification; **(b)** object detection; and **(c)** weakly supervised object detection. (A), (B), (C), and (D) are as defined in Section 4. The models for each subplot are, in the color order displayed, as follows: **(a)** MobileNet, DenseNet, ResNet; **(b)** YOLOv5, RetinaNet, SSD, Faster-RCNN; **(c)** Wetectron, C-MIL, WSOD2, PCL.

their jersey numbers even under such conditions, given that it includes many training examples of distorted and partially occluded jersey numbers. While a potentially negative societal impact of such a model would be that this would relieve coaches of this responsibility, coaches could instead focus on analyzing the personnel information.

Another useful application of a model trained on our dataset is that high school and college sports coaches may be able to track the movements of their players using game footage. While many professional sports teams use microchips to track the movements of their players, high school and college teams often do not have access to this level of technology. Our model would be able to locate players using their jersey numbers at each frame and be able to chart their movements, which is useful information that enables complex sports analysis.

6.2 Self-Driving Cars

In order to stay within speed limits, self-driving cars rely on internal vision systems to detect and read speed limits on the road [53, 54], which are most often printed on signs or painted directly onto the road. The consequences of a self-driving car being unable to read the speed limit may be dire, especially in an area where the speed limit warnings are few and far between.

Some speed limit signs may be old and worn-out, causing the digits to be faded or otherwise distorted, while speed limits painted on the road may be chipped and eroded. A self-driving car may not detect such speed limit postings, especially since it may be traveling past them at high speeds. By training its

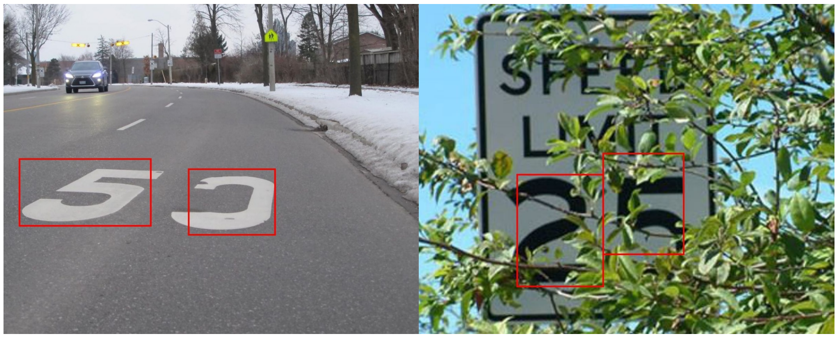

Fig. 5: Sample detections on worn-out and partially occluded speed limit postings using Faster-RCNN trained on CaltechFN.

vision systems on our dataset, which contains many examples of distorted digits, a self-driving car may be better equipped to read speed limits even under imperfect conditions. In other cases, speed limit signs and speed limits painted on the road may be partially occluded by other cars, pedestrians, or other obstacles. A self-driving car trained on our dataset, which contains many examples of partially occluded digits, would be better able to read such speed limit postings. To demonstrate the viability of this application, we applied the Faster-RCNN model mentioned in Section 4 to two sample images containing distorted (worn out) and partially occluded speed limit postings. The bounding box predictions are shown in Fig. 5.

7 Discussion and Future Work

In this paper, we have introduced CaltechFN, a new dataset containing digits found on American football jerseys. This dataset is novel in its variability: each digit is distorted or partially occluded in a unique way such that current models have difficulty learning the representative features of each digit class. We queried the Google Image Search database to collect our images, deleted images with no identifiable digits, then utilized Amazon Mechanical Turk to create annotations for the digits in each image. Through our experiments with various image classification, object detection, and WSOD models, we have demonstrated that CaltechFN is indeed a more difficult benchmark than SVHN and that models trained on CaltechFN demonstrate high cross-dataset generalization. Furthermore, we recorded human performance benchmarks on the CaltechFN dataset using experiments with human volunteers. In doing so, we illustrated the existing gap between model performance and human performance on our dataset. With this dataset, we aim to introduce a new state-of-the-art benchmark that will be used to foster the development of novel computer vision models. We hope that innovations in computer vision research will allow future models to achieve human levels of performance on our dataset.

We believe that models that perform well on our dataset will be better equipped to perform real-world tasks. Two such real-world applications of our dataset were described in Section 5. Models trained on our dataset could be used to automate the process of charting the personnel on the field at a given moment in a game. Also, self-driving cars can train its vision systems on our dataset to be better equipped to read speed limit postings under imperfect conditions.

7.1 Further Labeling

The images in the CaltechFN dataset contain rich information yet to be annotated, beyond the existing digit annotations that are the focus of this paper. Thus, we believe that the primary future direction of the CaltechFN dataset is to expand upon the existing annotations for the following applications:

Scene Recognition. Scene recognition refers to the computer vision task of identifying the context of a scene within an image. There is extensive research being conducted on the development of models to improve performance in this task [55–57]. While there do exist several large datasets for scene recognition, the task remains challenging and largely unsolved. These challenges have been attributed to class overlaps and high variance within classes [58]. Class overlaps occur when there are several classes that are not sufficiently distinct from each other, while high variance within classes means that classes have a wide variety of scenes attached to them. We believe that the unexploited qualities of our dataset may be able to address these challenges and improve model performance in the task of scene recognition. Within American football, each scene has a distinctive action being performed: tackling, throwing, catching, running, kicking, and so on. There is very little ambiguity between which of these actions is being performed in which scene. This lack of variance thus addresses the challenge of having class overlap—no two scene classes are the same and models should be better able to distinguish between the scene classes in this dataset.

Instance Segmentation. Another commonly studied computer vision task is that of instance segmentation [59–61]. This task is similar to object detection, with the difference being that the output consists of the set of pixels contained within the object rather than a rectangular box that bounds the object. Each object in an image segmentation dataset is annotated with its exact pixel-level boundary, rather than a simple rectangular bounding box. The images in our dataset contain many distinct objects that can be delineated in this way, including players, jerseys, helmets, and balls. We see that many of these objects are shaped in the form of some part of a human being. Thus, by adding pixel-level boundary annotations for these objects, our dataset could be used as a benchmark for the advancement of human detection and tracking techniques.

Acknowledgments. We would like to thank Elijah Cole, Yisong Yue, and A. E. Hosoi for their valuable comments and suggestions. We would also like to thank the members of our community who participated in our human performance benchmark experiments. Funding for all crowdsource worker tasks was provided by the George W. Housner Fund.

References

1. Y. LeCun, C. Cortes, and C. Burges, "THE MNIST DATABASE of handwritten digits," 1999. [Online]. Available: http://yann.lecun.com/exdb/mnist/. [Accessed: 15-May-2022].
2. D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
3. E. Kussul and T. Baidyk, "Improved method of handwritten digit recognition tested on MNIST database," *Image and Vision Computing*, vol. 22, no. 12, pp. 971–981, 2004.
4. S. H. Hasanpour, M. Rouhani, M. Fayyaz, and M. Sabokrou, "Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures," *arXiv*, 2016.
5. D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification," *2011 International Conference on Document Analysis and Recognition*, 2011.
6. C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019.
7. Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," *npj Computational Materials*, vol. 4, no. 1, 2018.
8. Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, Dec. 2011.
9. I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks," *arXiv*, Dec. 2013.
10. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, Jun. 2014.
11. A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR (Poster)*, 2016.
12. S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast AutoAugment," *Advances in Neural Information Processing Systems*, 2019.
13. S. N. Gowda and C. Yuan, "ColorNet: Investigating the importance of color spaces for image classification," *ACCV 2018*, pp. 581–596, 2019.
14. T. DeVries and G. W. Taylor, "Improved Regularization of Convolutional Neural Networks with Cutout," *arXiv*, 2017.
15. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
16. P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware Minimization for Efficiently Improving Generalization," *International Conference on Learning Representations*, 2021.
17. N. H. Phong and B. Ribeiro, "Rethinking Recurrent Neural Networks and Other Improvements for Image Classification," *arXiv*, 2020.

18. P. Rahimian and L. Toka, "Optical tracking in Team Sports," *Journal of Quantitative Analysis in Sports*, vol. 18, no. 1, pp. 35–57, 2022.

19. T. B. Moeslund, G. Thomas, and A. Hilton, *Computer Vision in Sports*. Cham: Springer International Publishing, 2015.

20. D. Bhargavi, E. P. Coyotl, and S. Gholami, "Knock, knock. Who's there? – Identifying football player jersey numbers with synthetic data," *arXiv*, 2022.

21. I. Atmosukarto, B. Ghanem, S. Ahuja, K. Muthuswamy, and N. Ahuja, "Automatic Recognition of Offensive Team Formation in American Football Plays," *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

22. E. Chong, A. M. Familiar, and W. M. Shim, "Reconstructing representations of dynamic visual objects in early visual cortex," *Proceedings of the National Academy of Sciences*, vol. 113, no. 5, pp. 1453–1458, 2015.

23. P. Kok and F. P. de Lange, "Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex," *Current Biology*, vol. 24, no. 13, pp. 1531–1535, 2014.

24. G. Bosco, S. Delle Monache, S. Gravano, I. Indovina, B. La Scaleia, V. Maffei, M. Zago, and F. Lacquaniti, "Filling gaps in visual motion for target capture," *Frontiers in Integrative Neuroscience*, vol. 9, 2015.

25. Y. Revina and G. W. Maus, "Stronger perceptual filling-in of spatiotemporal information in the blind spot compared with artificial gaps," *Journal of Vision*, vol. 20, no. 4, p. 20, 2020.

26. B. Chandler and E. Mingolla, "Mitigation of Effects of Occlusion on Object Recognition with Deep Neural Networks through Low-Level Image Completion," *Computational Intelligence and Neuroscience*, vol. 2016, pp. 1–15, 2016.

27. C. Ning, L. Menglu, Y. Hao, S. Xueping, and L. Yunhong, "Survey of pedestrian detection with occlusion," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 577–587, 2020.

28. D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," *ECCV 2012*, pp. 340–353, 2012.

29. J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

30. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

31. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," 2010.

32. H. Kusetogullari, A. Yavariabdi, A. Cheddad, H. Grahn, and J. Hall, "ARDIS: a Swedish historical handwritten digit dataset," *Neural Computing and Applications*, vol. 32, no. 21, pp. 16505–16518, 2019.

33. T. de Campos, B. R. Babu, and M. Varma, "Character Recognition in Natural Images," *VISAPP 2009 - Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, vol. 2, 2009.

34. N. Nayef, M. M. Luqman, S. Prum, S. Eskenazi, J. Chazalon, and J.-M. Ogier, "SmartDoc-QA: A dataset for quality assessment of smartphone captured document images - single and multiple distortions," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015.
35. S. Gerke, K. Müller, and R. Schäfer, "Soccer Jersey Number Recognition Using Convolutional Neural Networks," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.
36. H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware Pyramid Reconstruction for Alignment-free Occluded Person Re-identification," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
37. X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," *2013 IEEE International Conference on Computer Vision*, 2013.
38. Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," *IEEE Winter Conference on Applications of Computer Vision*, 2014.
39. K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3388–3415, 2021.
40. I. Voxel51, "Voxel51 // developer tools for ML," *Voxel51 // Developer tools for ML*. [Online]. Available: https://voxel51.com/. [Accessed: 08-Jun-2022].
41. A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.
42. Amazon Mechanical Turk. [Online]. Available: https://www.mturk.com/mturk/welcome. [Accessed: 16-May-2022].
43. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv*, 2017.
44. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
45. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv*, 2020.
46. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for Dense Object Detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
47. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *European Conference on Computer Vision (ECCV) 2016*, 2016.
48. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.
49. Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

50. F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
51. Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning Bottom-up and Top-down Objectness Distillation for Weakly-supervised Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
52. P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal Cluster Learning for Weakly Supervised Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition 2017*, 2017.
53. N. Kanagaraj, D. Hicks, A. Goyal, S. Tiwari, and G. Singh, "Deep learning using computer vision in self driving cars for lane and traffic sign detection," *International Journal of System Assurance Engineering and Management*, vol. 12, no. 6, pp. 1011–1025, 2021.
54. W. A. Farag, "Recognition of traffic signs by convolutional neural nets for self-driving vehicles," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 2018.
55. L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: objects, scales and dataset bias," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
56. G. Chen, X. Song, B. Wang, and S. Jiang, "See More for Scene: Pairwise Consistency Learning for Scene Classification," *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.
57. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.
58. A. Matei, A. Glavan, and E. Talavera, "Deep learning for scene recognition from visual data: a survey," *arXiv*, 2020.
59. D. A. Ganea, B. Boom, and R. Poppe, "Incremental Few-Shot Instance Segmentation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
60. Y. Wang, Z. Xu, H. Shen, B. Cheng, and L. Yang, "CenterMask: Single Shot Instance Segmentation with Point Representation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
61. E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "PolarMask: Single Shot Instance Segmentation With Polar Representation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.