Blind Users Accessing Their Training Images in Teachable Object Recognizers

JONGGI HONG, Smith-Kettlewell Eye Research Institute, United States
JAINA GANDHI, University of Maryland, College Park, United States
ERNEST ESSUAH MENSAH, University of Maryland, College Park, United States
FARNAZ ZAMIRI ZERAATI, University of Maryland, College Park, United States
EBRIMA HADDY JARJUE, University of Maryland, College Park, United States
KYUNGJUN LEE, University of Maryland, College Park, United States
HERNISA KACORRI, University of Maryland, College Park, United States

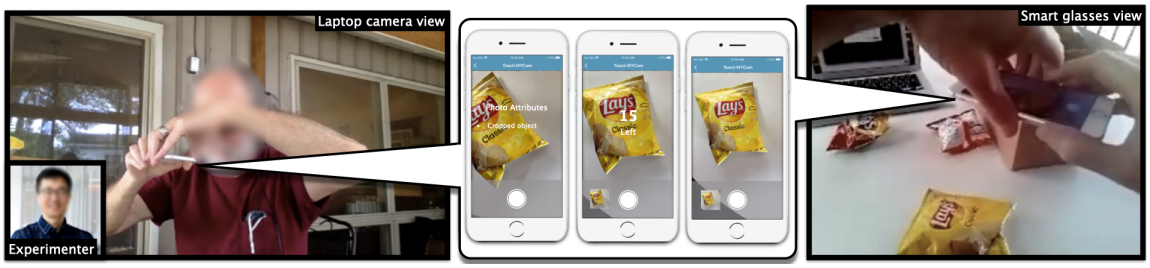


Fig. 1. A blind participant in our study training the MYCam app in their homes to recognize Lays with real-time descriptors. A dual video conferencing captures participant's activities via a laptop camera and smart glasses worn by the participant.

Teachable object recognizers provide a solution for a very practical need for blind people – instance level object recognition. They assume one can visually inspect the photos they provide for training, a critical and inaccessible step for those who are blind. In this work, we engineer data descriptors that address this challenge. They indicate in real time whether the object in the photo is cropped or too small, a hand is included, the photos is blurred, and how much photos vary from each other. Our descriptors are built into open source testbed iOS app, called MYCam. In a remote user study in (N = 12) blind participants' homes, we show how descriptors, even when error-prone, support experimentation and have a positive impact in the quality of training set that can translate to model performance though this gain is not uniform. Participants found the app simple to use indicating that they could effectively train it and that the descriptors were useful. However, many found the training being tedious, opening discussions around the need for balance between information, time, and cognitive load.

Additional Key Words and Phrases: blind, visual impairment, object recognition, machine teaching, participatory machine learning

ACM Reference Format:
Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Haddy Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22), October 23–26, 2022, Athens, Greece.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3517428.3544824

1 INTRODUCTION

Echoing the end-user programming paradigm, the idea of having end-users consciously provide training examples in AI-infused applications has recently gained traction along with advances in neural networks. By leveraging prior work in transfer, meta, and few-shot learning (*e.g.*, [7, 11, 70, 76, 80]), we are now able to build *teachable* applications, where end-users can train models of their own. These applications facilitate personalization as they promise a better fit for real-world scenarios by significantly constraining the machine learning task to a specific user and their environment. Thus, it is no surprise to see early mentions of the term "teachable" in accessibility research (*e.g.* [55]), where data is sparse and there is a high diversity even for a given disability [32, 33]. A more recent example (and the focus of this work) is teachable object recognizers [3, 34, 41, 42, 47, 68], where blind users train their camera-equipped devices such as mobile phones to recognize everyday objects by providing a few photos as training examples.

Why is it important to access one's training examples? In teachable applications such as teachable object recognizers, users are called to interact with the machine learning model and improve its performance by accessing and controlling their examples [83]. Personalization is often the ultimate goal. However, the interactive nature of these applications can also help people uncover basic machine learning concepts and gain familiarity with AI (*e.g.*, [14, 17, 25, 27, 56]). Thus, they can also contribute to the larger goal of "*making the process of teaching machines easy, fast and above all, universally accessible*" [65]. An underlying assumption for both improving a model and uncovering concepts via experimentation is that users can inspect their data and iterate the training and testing. By doing so, they could build an intuition about what works and what doesn't and perhaps why. However, this assumption does not often hold for assistive teachable applications. Inspecting training examples typically requires similar skills to those the technology aims to fulfill [20, 31]; thus, it is often inaccessible. For example, *teachable object recognizers, where users teach the model to recognize objects on their behalf, assume that they can see the training images they are providing, which is almost never the case with blind users.* Sure enough, blind participants in prior studies with this technology wanted to know more about their training examples [31] with one of them stating "*the most challenging and most fun is training the person*".

Existing approaches for real-time '*alt text*' for individual images and '*scene description*' for a series of images are not suitable for this task; they do not capture fine-grained differences across otherwise similar images (*e.g.*, see Figure 1). In this paper, we explore this **challenge of accessing one's training data**. Within the context of teachable object recognizers for the blind, we study the potential and limitations of real-time '*data descriptors*' that can capture users' training examples with photo- and set-level attributes. Specifically, we investigate whether these descriptors could be derived from visual attributes used to code training photos from sighted (*e.g.*, [26, 27]) and blind (*e.g.*, [34, 41]) people. To this end, we engineer photo-level descriptors that communicate to the user in real-time information about the photo they just took such as blurriness, presence of their hand, object visibility, and framing. We also engineer set-level descriptors that communicate information one would get from glancing over a group of training photos such as variation in object background, distance, and perspective; all factors that can affect model performance.

Through a remote user study with 12 blind participants (with the setup shown in Figure 1), we demonstrate that our data descriptors support blind users in reducing photos with cropped objects in their training sets and increase variation. Many participants chose to iterate after inspecting their training sets and reflected by improving many photo attributes, which resulted in models that generalize better to photos from others, even though they reduced variation in background. Aligned with prior studies, we also observe challenges among participants in crafting good testing examples that could further promote experimentation. Still, their models perform better when tested on their own photos compared to both aggregated test sets from all 12 blind participants in our remote study and from 9 blind participants

in an in-lab study [41]. Observations from our analysis of the photos and model performance are also confirmed by participants' subjective feedback. Responses support the potential of descriptors, with blind participants indicating that they were able to tell their meaning by looking at relative changes in values and finding them useful. However, errors in descriptors affected the reliability of the app for some. More so, some considered training being tedious referring both to time and cognitive load (*e.g.*, optimize for multiple variables). Many made design recommendations that could further improve the effectiveness of the descriptors and the training process.

To the best of our knowledge, this is the first work to propose non-visual access to training data and to provide empirical results with blind participants on automatically estimating and incorporating descriptors for data inspection in teachable computer vision applications. Our analysis focuses on object recognizers, where 'learning to train' is deemed as one of the main challenges among blind users [31, 34]. However, we see how the underlying methods for extracting meaningful instance- and set-level descriptors can be adopted for other teachable applications both in assistive and informal AI learning contexts. Perhaps, they can also serve towards more accessible approaches for explainable AI interfaces, where there is an underlying assumption on people's ability to visually inspect explanations [62, 66, 69].

2 RELATED WORK

There is a rich literature exploring how computer vision can benefit people with disabilities (*e.g.*, [10, 13, 30, 36, 46, 58, 71]). This is especially the case with assistive technologies for the blind, where computer vision is employed on smartphones (*e.g.*, [2, 24, 38, 60, 78, 82]), smart glasses (*e.g.*, [18, 43, 67, 81]), and smart suitcases (*e.g.*, [23, 35]). A common challenge we share with prior work is that aiming the camera and inspecting recognition errors typically requires similar skills to those the technology aims to fulfill (*i.e.*, sight), even though the majority of prior work employs AI-infused systems pre-trained by engineers, not fine-tuned by the end-user. Thus, it is not a surprise to find that recognition errors affect blind users' experience [51]; sometimes, to a degree where it can not be corrected even by human clarification [61]. In fact, blind users may depend on the recognition especially when it is difficult to verify its predictions. They may overtrust the predictions even when they know they can be error-prone [41, 45] though, errors are especially non-acceptable when they can adversely affect interactions with others [1, 43]. Aligned with prior efforts aiming to support users' recovery from errors [6, 28, 52], we explore how to make training and resolving errors in teachable object recognizers more accessible to blind users. Below, we focus on prior work that closely relates to ours and contrast it to our study.

2.1 Teachable Object Recognizers

Looking at prior work on teachable object recognizers, we see diversity in research aims. Some, similar to this work, focus on the blind community. They explore the potential of teachable object recognizers as an assistive technology for blind users [34, 68], build feedback mechanisms for better camera aiming [3, 41], and collect benchmarking datasets for evaluating approaches in transfer learning and meta learning [72]. Our work is orthogonal and highly complementary to these efforts – our shared goal is to improve blind users' experience with teachable object recognizers.

We also see studies involving sighted people both adults (*e.g.*, [27]) and children (*e.g.*, [17, 75]). They aim to better understand the potential of teachable machines for enabling non-experts to uncover basic machine learning concepts as well as better understand common AI misconceptions they may have. Insights from these studies are very informative for our efforts in making the *'learning to train'* challenge more accessible to blind adults and perhaps in the future to blind children that may want to participate in similar informal learning activities as in Dwivedi *et al.* [17].

Table 1 provides a more detailed overview from a sample of these prior studies over the past five years (2017-2021) with the number of participants being typically smaller for in-person studies with blind people and sighted children. As

Table 1. Characteristics of related studies on teachable object recognizers juxtaposed with ours.

		[34]	[68]	[41]	[27]	[3]	[75]	[17]	[72]	This study
Participants	Blind Adults	8	14	9		10			52	12
	Sighted Adults	2	10	2	100					
	Sighted Children						6	14		
Setting	Real-world	•			•	•		•	•	•
	Controlled	•	•	•		•	•	•		
Input	Photo	•		•	•	•	•	•		•
	Video		•						•	
Tasks	Train	•		•	•	•	•	•	•	•
	Test	•	•	•	•	•	•	•	•	•
	Iterate				•		•	•		•
Access	Framing			•		•				
	Review				•		•	•		•

the performance of teachable object recognizers and users' behavior in taking photos can be affected by environmental factors such as background, light condition, and selection of objects, many studies collected inputs from participants' environments to incorporate these factors [3, 27, 34, 72]. We also opted for this approach in our study; the study was conducted in the homes of blind participants while we control for factors such as study procedure and object stimuli.

The majority of prior work on teachable object recognizers facilitates training through photos [3, 14, 27, 34, 41], except for one [72], where blind users are called to use short videos. In our study, we also used photos so that the outcomes of our study could be applicable (and comparable) to the majority of existing approaches. More so, collecting videos may increase the burden on the user, especially when they are given several instructions and tasks to do [72] as in the case of our study. In addition, video-based assistive technology can pose a greater privacy risk for blind users [5] as it is more likely to capture unwanted objects and unnecessary information in a video. Perhaps, live photos [53], could be the middle ground between the two. We further reflect on the potential of this approach in the discussion section.

In their early explorations, Kacorri *et al.* [34] highlighted some of the main challenges that blind users may face when training a teachable object recognizer and testing its performance. They revolve around camera framing (*i.e.*, adjusting the distance between the camera and an object and centering the object), capturing the side of the object with the most distinctive visual features (*i.e.*, product logos), and reviewing the training photos after taking them (*i.e.*, quality and characteristics). Lee *et al.* [41] and Ahmetovic *et al.* [3] aimed to resolve the camera framing challenge by developing real-time audio/haptic feedback that helps blind users estimate the proper distance and position of the object in the image frame [3, 41]. However, the challenge of reviewing photos for iteration has not been addressed yet.

Typically, studies included a training and testing step for exploring participants' interactions with the teachable interfaces. Very few of them [17, 27], though, allowed people to reflect and iterate giving them access to their training data for review. We believe iteration is a critical step for understanding the potential of *descriptors* for making the review process more accessible. Thus, in our study, we also provide blind participants with an opportunity to reflect and the option to iterate after reviewing their images with the descriptors. After all, our goal is to examine how data descriptors that provide non-visual access to training photos, either individually or as a set, can be helpful for blind users during the iterative process of training and testing, as well as how blind users may interact with them.

3 MYCAM: A TESTBED TEACHABLE OBJECT RECOGNIZER WITH DESCRIPTORS

To explore the potential and limitations of real-time *descriptors* derived from visual attributes for accessing one's training data, we build MYCam. MYCam serves as a testbed for deploying descriptors in a teachable object recognizer. In the background, it sends users' photos to a server, where an image recognition model is being fine-tuned by the user. While privacy is one of the promises of teachable object recognizers [31] (*i.e.*, by processing photos entirely on the user's mobile device), we find that the state-of-the-art on-device training is not there yet. As a screen-reader accessible iOS mobile app, MYCam enables remote studies with blind participants. This was critical for us; due to the pandemic, we had to move our study from the lab to blind participants' homes. By open sourcing both the MYCam app (available at https://iamlabumd.github.io/MYCam-Mobile/) and our proof-of-concept implementation of the descriptors (available at https://iamlabumd.github.io/MYCam-Server/), we are hoping that others can contribute to further advance this work.

3.1 Design Rationales

DR1: Prioritize Blind Users. Both the form factor and interaction modalities of MYCam are informed by prior work with blind users and teachable object recognizers as well as broader real-world object recognition applications. We opted for an iOS app since prior work in the United States, the location of our study participants, suggests that blind smartphone users overwhelmingly favor the iPhone [50] though the actual numbers may be changing these past years [49]. When users open the app, they enter the main screen (Figure 2a), which shows a camera preview. We opted for the default camera app in iOS maximizing both compatibility with VoiceOver and user familiarity with it. The recognition mode for MYCam was modeled after existing real-world applications, such as Seeing AI [63], where users

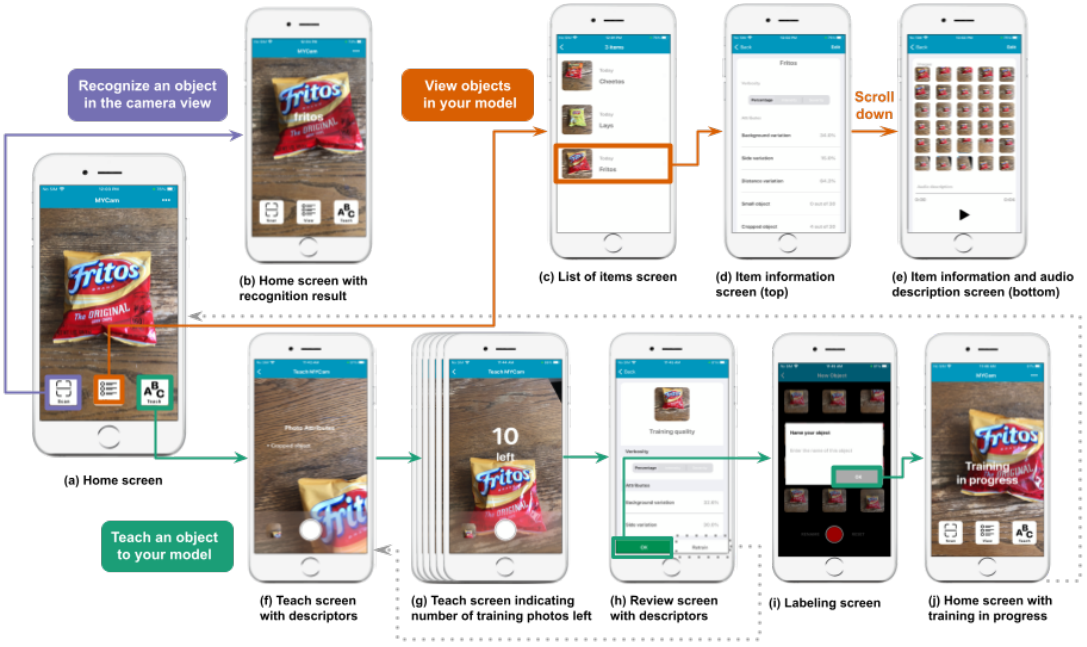


Fig. 2. The user flow of MyCam. MyCam has three main parts: Recognizing an object in the camera view (purple thread), reviewing and editing the information of the objects (red thread), and teaching an object to the model (green thread).

can immediately ask the app to recognize what is captured by the camera with a double-tap; the *Scan* button is activated by default. In this case, the app takes a photo, sends it to the personalized object recognition model in the server, and indicates the predicted label both via speech and visually (Figure 2b). To mitigate potential errors that can't be verified non-visually, the app says *"Don't know"* when uncertain (approach for uncertainty is discussed in Implementation).

DR2: Simplify the Machine Teaching Flow. Users can add a new object to the recognition model via the *Teach* button on the Home screen. The app displays the (rear-facing) camera preview with the shutter button at the bottom center and a thumbnail image of the last photo in the lower-left corner (Figure 2f). Users are asked to take 30 photos with the count indicated in real-time via speech and visually (Figure 2g); in Kacorri *et al.* [34], blind participants indicated that they would like to obtain feedback from the camera on the number of photos taken. The number of training examples (*i.e.*, 30) is also informed by the same study [34] with blind participants spending on average 65 seconds (*SD*=35.2) to take 30 photos and often providing variation in their training examples. More so, k-shot learning results in the literature are often reported for $k = 1$, 5, or 20. Thus, 30 examples could allow for bootstrap estimates for future comparisons in this field. As discussed in Related Work, the majority of prior work in teachable object recognizers opt for photos rather than videos – we followed this approach in hope that photos provide blind users with more control over their training examples in terms of both conscious variation incorporated and privacy concerns mitigated (*e.g.*, presence of their hand, bystanders, or surroundings in the camera frame). In the one study where videos were used, blind users had to be explicitly trained to record videos and follow specific filming techniques [72]. Given the emphasis of this study on the descriptors, we decided to simplify the machine teaching and not require any explicit training steps for the users. After the training examples, a dialogue box with a text field shows up prompting users to enter the name of the object (Figure 2i). More so, in this screen users can opt to add an audio description. (Both object name and description can be edited at a later time, as shown in Figures 2d and 2e.) Once this step is completed, the app notifies the user with a "Training in progress" message (Figure 2j). At this point *Scan* and *Teach* buttons are made inactive. They are activated once training on the server is complete and the user is notified.

DR3: Enable Access to Training Data with Descriptors. Some of the main concerns of blind participants about teachable object recognizers in Kacorri *et al.* [34] were: "*knowing whether the photos were good, knowing the area of a package where the label or distinguishing information resides,.., and deciding on the distance between the object and camera lens.*" We observe that this information wanted by the participants can be provided both at a photo level and at a higher level across a set of photos. Thus, we devise two types of descriptors, shown in Table 2. These are all derived from visual attributes used to code training photos from sighted (*e.g.*, [26, 27]) and blind (*e.g.*, [34, 41]) people. Photo-level descriptors are binary, they indicate whether the object is too small or partially included in the frame (cropped), whether the photo is blurred, and if user's hand is included in the frame. Set-level descriptors are indicated as a percentage. They draw from parallels to how humans recognize objects independent of size, viewpoint, and location [54].

As shown in Figure 2f, users access the photo-level descriptors after every photo that they take so that they can identify problems in the photo (*e.g.*, object being cropped) right away; since this gets repetitive, a photo-level descriptor is communicated only when true. Users can access this detailed information also later, when reviewing their trained objects (Figure 2e). Photo-level descriptors are also provided in aggregate together with the set-level descriptors (*e.g.*, photo blurred in 50% of the training examples for an object). Users can access these aggregates along with the set-level descriptors at the end of a training session (Figure 2h), where they are called to select either *OK* to proceed or *Retrain* to retake the photos from scratch. Both photo-level and set-level descriptors can be accessed at a later time when reviewing and editing trained objects, as shown in Figure 2d.

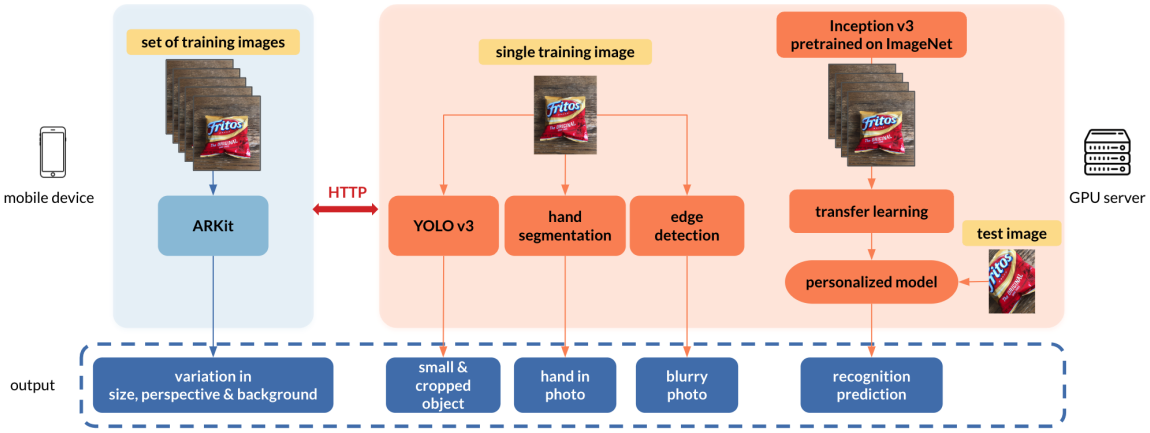


Fig. 3. The architecture of the MYCam system indicating approaches for estimating the descriptors and recognizing the object.

3.2 Implementation

We built the MYCam testbed on Apple iPhone 8 with the object recognition models and descriptor estimators running on our server on an NVIDIA GeForce GTX 1080 Ti GPU; the two communicate through HTTP. The architecture of the system, indicating how both descriptors and recognition predictions are obtained, is illustrated in Figure 3. The estimation of the descriptors in the current implementation of MYCam is error-prone; our approaches merely serve as proof of concept. Prior to making these approaches more robust, we wanted to examine whether blind users can leverage such descriptors in the first place for accessing their training data and experimenting with the model.

3.2.1 Descriptors. In all previous studies that informed our descriptors, researchers coded the attributes of photos manually through visual inspection of the photos from participants. Given that this is a time-consuming process, methods like Wizard of Oz do not deem appropriate in this early exploration of descriptors for facilitating accessible non-visual experimentation. Thus, we opt for methods that attempt to automatically estimate them, even though, developing techniques for more accurate estimations is beyond the focus of this paper and is briefly discussed in Section 6. Specifically, we employ state-of-the-art computer vision techniques such as world tracking in ARKit, a YOLOv3 object detection model [57], and hand segmentation models [42] to estimate the descriptors. To speed up the calculations for real-time interactions, the object detection, hand segmentation, and edge detection run on our server.

Table 2. Photo-level and set-level descriptors. The descriptors are informed by prior studies with sighted and blind people who have no machine learning expertise looking at the way they synthesize their data for training [17, 26, 27, 34, 41].

Photo-level descriptors	
Small object	The bounding box of the object is smaller than 1/8 (12.5%) of the image.
Cropped object	The object is partially included in the image.
Blurry photo	The photo is too blurry to recognize textures or texts.
Hand in photo	A user's hand is visible in the image.
Set-level descriptors	
Variation in size	A set of images shows objects with different sizes.
Variation in perspective	A set of images shows different sides of objects.
Variation in background	A set of images show backgrounds with different textures or items.

- **Small object**: Given a bounding box of an object in an image from YOLOv3 [57], the object is considered too small if the size of its bounding box is smaller than 1/8 (12.5%) of the image.
- **Cropped object**: If the YOLOv3 [57] bounding box is at the edge of the image, the object is considered cropped.
- **Blurry photo**: The original RGB image is converted to grayscale (pixels values range: 0-255). We use Laplacian edge detection [37] to produce an image with the edges in the grayscale image. In this last image, we then calculate variance in pixel values to quantify blurriness. If the variance is lower than a threshold, the photo is considered blurry. In this study, we set the threshold at 3.0; we found it classifies the blurriness most accurately when tested on photos collected in a prior study with blind participants [41].
- **Hand in photo**: The server detects the pixels from a hand via a hand segmentation model that has been previously tested with blind participants [42]. If the proportion of pixels of a hand(s) in an image is greater than a threshold, it considers the photo to show a hand. The threshold is 0.3%, which detected photos with hands most accurately when tested with the photos collected in a prior study by Lee *et al.* [41]
- **Variation in size**: When users take a photo, we detect the position of the smartphone with ARKit. As the size of the object depends on the distance between the phone and the object, we used the standard deviation of the differences between the phone positions (SD_{pos}) to measure the variation in size indirectly. We set the maximum value of the variation as 0.15 (SD_{max}) that we could observe with the photos collected by a sighted person in our research team through an internal test. The app presents the variation in size as percentage ($SD_{pos}/SD_{max} * 100$).
- **Variation in perspective**: We detect the sides of an object using ARKit. For this, we pre-trained the 3D object detection model in ARKit with the three object stimuli in our study. The model provides an enclosing bounding box of an object with six sides in 3D space when it detects the object regardless of the object shape. The model finds the main side of the bounding box based on the object orientation. We calculate the variation in perspective based on the number of object sides shown in a training set. For consistency with other descriptors, we present the variation in perspective as a percentage with scaling ($n * 15\%$ where n is the number of sides in photos).
- **Variation in background**: Assuming that the backgrounds captured in photos can vary as a user moves the camera to different places or changes its orientation, we used the location and orientation of the camera to measure the variation in the background. We calculate the standard deviations of differences in both orientation (using 1-cosine similarity) and the location of the camera in the 3D coordinate system in ARKit. The greater value of the two standard deviations is selected as a variation in background. Like variation in size, we set the maximum value as 0.15 through an internal test. We present the variation in background as a percentage.

3.2.2 Object Recognition Model. The base model for object recognition is Inception V3 pre-trained on ImageNet [15]. When users train the app, it fine-tunes the last layer of the base model using transfer learning with photos taken by the users. The transfer learning works with a gradient descent algorithm with 500 iterations and a 0.01 learning rate. The training takes around 80 seconds with 90 photos of three objects. When users recognize an object with a personalized model, the time from taking a photo to notifying the recognition result is around 100 milliseconds. To make the model distinguish the objects in a user's training set and tell the difference from other objects that it has not been trained on, we employed an approach of quantifying the confidence level of the discriminability based on the entropy of confidence scores [79]. Specifically, when the entropy value is greater than 2.0 or the confidence score is lower than 0.4, the app says *"Don't know"* in synthesized speech instead of the label predicted by the model. We decided the thresholds of the entropy and confidence score through internal tests such that the app could differentiate the three objects for the user study in the Section 4 from other items (*e.g.*, pen, keyboard, mouse, keys) with the thresholds most accurately.

4 USER STUDY

To explore the potential and limitations of descriptors in the context of a teachable object recognizer, we conducted a remote user study with blind participants. The study took place in participants' homes to minimize safety concerns during the COVID-19 pandemic. The study was approved by the Institutional Review Board at the University of Maryland, College Park (IRB #1255427-1). In designing this remote study, we came across many challenges, including how to provide remote guidance and observe participants' interactions with MYCam and their objects. We quickly found that having just the third-person camera view from the laptop was not enough. Thus, as shown in Figure 1, we added a first-person view with smart glasses. We iterated via several pilot tests that involved blind and sighted researchers in our team to anticipate the logistics (*i.e.*, study equipment delivery) and communication methods (*i.e.*, laptop and smart glasses) required for this remote study. Lessons learned from accessing blind participants' interactions via smart glasses (with this study serving as part of a larger case study) are discussed in depth in Lee *et al.* [40].

4.1 Participants

We recruited 12 blind participants (6 women, 6 men, 0 nonbinary) from campus email lists and local organizations. As shown in Table 3, their ages ranged from 32 to 70 ($M = 54.3, SD = 15.2$). Three participants reported being totally blind, five having some light perception, and four being legally blind. P1 and P2 reported having an "*auditory processing disorder*" and difficulty in hearing "*very high sound*", respectively. All participants reported using smartphones several times a day and taking a photo or recording a video at least once a month. As for their familiarity with machine learning, two participants reported being somewhat familiar, eight being slightly familiar, and two being not familiar at all—we used a 4-point scale for this question: (1) not familiar at all (have never heard of machine learning), (2) slightly familiar (have heard of it but don't know what it does), (3) somewhat familiar (have a broad understanding of what it is and what it does), (4) extremely familiar (have extensive knowledge on machine learning). While all participants had experience taking photos before, many indicated that they had challenges related to image framing (9), focusing (2), holding a camera steadily (2), and controlling the lighting (2). Many participants indicated prior experience with other camera-based assistive mobile applications such as Aira [4], Be My Eyes [8], Google Lookout [21], Microsoft Seeing AI [63], Mediate Labs Supersense [48], Super Lidar [29], and Voice Dream Scanner [16].

Table 3. Participants' demographics and experience with machine learning, photo taking, and camera-based assistive apps.

ID	Age	Gender	Level of vision	Onset	Machine learning	Photo taking	Experience with assistive apps
P1	39	Woman	Light perception	Birth	Not familiar at all	Once a day	Aira, Be My Eyes, Seeing AI
P2	67	Man	Legally blind	55	Slightly familiar	Once a month	Seeing AI
P3	62	Woman	Totally blind	Birth	Somewhat familiar	Several times a month	Seeing AI, Be My Eyes
P4	32	Man	Legally blind	20	Slightly familiar	Several times a day	None
P5	66	Man	Light perception	46	Slightly familiar	Once a week	Seeing AI, Supersense, Super Lidar
P6	61	Man	Light perception	41	Somewhat familiar	Several times a week	Seeing AI
P7	70	Man	Legally blind	Birth	Slightly familiar	Several times a week	None
P8	50	Woman	Legally blind	45	Slightly familiar	Several times a week	Seeing AI
P9	69	Woman	Totally blind	55	Not familiar at all	Several times a day	VD Scanner, Be My Eyes, Seeing AI
P10	66	Woman	Light perception	Birth	Slightly familiar	Several times a week	None
P11	33	Woman	Light perception	Birth	Slightly familiar	Once a month	Seeing AI, VD Scanner
P12	36	Man	Totally blind	Birth	Slightly familiar	Several times a day	Seeing AI, VD Scanner, Lookout

4.2 Procedure

Participants communicated with the experimenter remotely via dual Zoom video conferencing [84] connected both via a laptop and a pair of Vuzix Blade smart glasses [77] that we delivered prior to their study sessions (see Lee *et al.* [40]). At the beginning of the study, we briefly explained the concept of a teachable object recognizer. Here, we provided a minimal description of how to take photos to train or test the app to mitigate priming in photo-taking strategies for training and testing an object recognizer. The description given at the beginning of the study reads as follows:

> "The idea behind the app is that you can teach it to recognize objects by giving it a few photos of them, their names, and if you wish, audio descriptions. Once you've trained the app and it has them in its memory, you can point it to an object, take a photo, and it will tell you what it is. You can always go back and manage its memory."

Then, participants were asked to perform three tasks: (1) train the app with their own photos and labels of three snacks that served as object stimuli shown in Figure 4, (2) use the app again to recognize those objects later *i.e.* to test the performance of the app, and (3) review and edit the information of the already trained objects. For the first task, the order of objects for training was fully counterbalanced between participants. When participants trained the app with the first object, the experimenter provided step-by-step instructions on the MYCam user interface (*e.g.*, the position and functionality of buttons as well as the audio feedback that indicates the steps of training). Then, participants trained the app with the second and third objects and asked the experimenter for help when necessary. When participants were testing the app for the first time, the experimenter also gave detailed instructions on the MYCam interface for testing. After that, participants were free to test their models for as long as they wished (taking any number of photos). When reviewing their trained objects in the third task, participants could access both information related to the descriptors as well as their own object labels and any audio descriptions they may have recorded.

After reviewing a training set with the descriptors, participants decided whether they would collect the photos again or not for that object. We made retraining optional for two reasons: (1) to avoid collecting data from participants who are not motivated to experiment by retraining a model (as this could add a confounding factor in our analysis) and (2) to be able to contrast the attributes of training sets for who decided to retrain their models and those who did not.

Throughout the study, we encouraged participants to think out loud and to ask questions at any time. After each task, participants were asked to answer questions related to their experience with the descriptors and MYCam and questions captioning usability satisfaction [44]. All questions in this study were either open-ended or on a 5-point Likert scale (*i.e.*, strongly disagree, disagree, neutral, agree, strongly agree).

4.3 Object Stimuli

Accounting for blind people's need for recognizing objects with similar sizes, weights, and textures with fine-grained labels [34, 68], we selected three snacks, shown in Figure 4, with the same size, texture, and nearly identical weights for our user study. As prior work shows that end-users' strategies of collecting training photos are often inconsistent between objects [27], we expect that the choice of three similar objects allows us to observe blind people's teaching strategies in the context of fine-grained object recognition. With these snacks, we simulated a scenario in which a blind user interacts with the app to recognize different objects that the blind user may feel difficult to distinguish using only the tactile sensation. It was engineered to be a challenging scenario for machine learning models since these objects were similarly shaped and colored, had reflective surfaces, and were deformable. Unique and personal objects without logos or texts on them (*e.g.*, key, mug cup) can be potentially used with a teachable object recognizer and perhaps

Fig. 4. Object stimuli in the study simulating a challenging fine-grained classification task: Fritos, Cheetos, and Lays.

could be fit for a more realistic scenario. However, for this study, we included only commercial products to allow for comparison and replicability similar to prior studies regarding teachable object recognizers [34, 41, 68].

5 RESULTS

Participants spend on average 143.8 seconds ($SD = 72.4$) taking 30 photos of an object. Five out of 12 participants re-train the object recognizer after inspecting their training sets with descriptors. Examples of the photo-collection attempts and their annotated attributes (*i.e.*, ground-truth attributes annotated by a researcher through visual inspection) are shown

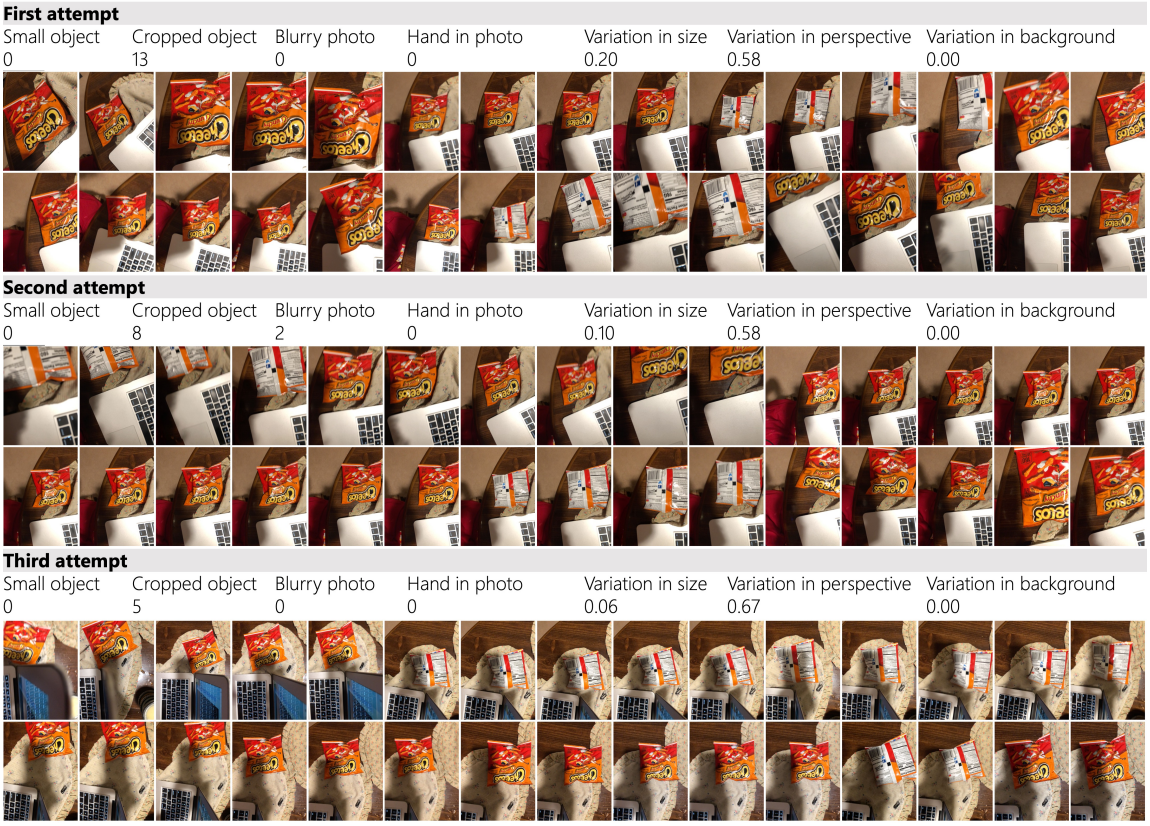


Fig. 5. Photos of Cheetos from P10 and manually annotated attributes to be compared with automatically estimated descriptors.

in Figure 5. Through the analysis of the participants' photos and the performance of the personalized object recognition models, we show how descriptors may relate to the participants' strategies for collecting training photos when they decided to retrain their models. We also show the impact of these changes in training photos on the performance of the models. We observe promising trends in the characteristics of photos (*i.e.*, adding more variations and reducing problematic photos) over time and iterations. Participants' subjective feedback also indicate that our descriptors can be a promising approach for providing access to one's training data in this context.

5.1 Correlation Between Estimated Descriptors and Annotated Attributes

We report the performance of our approach in estimating descriptors as it is a critical context for interpreting the remainder of the results. More so, it can provide a glimpse at future efforts for estimating such descriptors in a real-world context. Here we measure performance by computing the correlation between the estimated descriptors and annotated attributes. Given that prior work indicates high inter-rater agreement for the annotation of these attributes [27], we had a single researcher in our team performing this task. To quantify the variation of background and perspective, the researcher grouped the photos within a set based on their similarity in terms of background and object side. We used the groups to calculate the Shannon-Wiener Diversity Index [64], a measure of variation in background and perspective. The researcher also coded the photos with a cropped object, participants' hands, and blurriness. For the attributes related to the size of the object (*i.e.*, variation in size, small object), the researcher annotated the bounding boxes of the objects. The variation in size was considered as the standard deviation of the proportions that the bounding boxes occupy in photos. The proportions range from 0.0 (*i.e.*, the object is not captured) to 1.0 (*i.e.*, a bounding box covers the entire photo). A photo with a small object is defined as one having a bounding box smaller than 12.5% of the photo.

As shown in Figure 6, the correlation coefficients between estimated descriptors and annotated attributes ranged from 0.23 to 0.57, highlighting that this is a challenging task. The correlation for "small object" is not shown since only three of all photos had small objects that are not detected by our descriptor estimator. Even though we employed naive approaches for estimating the descriptors as a proof of concept, all pairs had positive correlations. This indicates that even with partial access there can be an opportunity for reflection and experimentation *i.e.*, if participants considered relative changes rather than absolute values. Below, we see some empirical evidence in support of this premise.

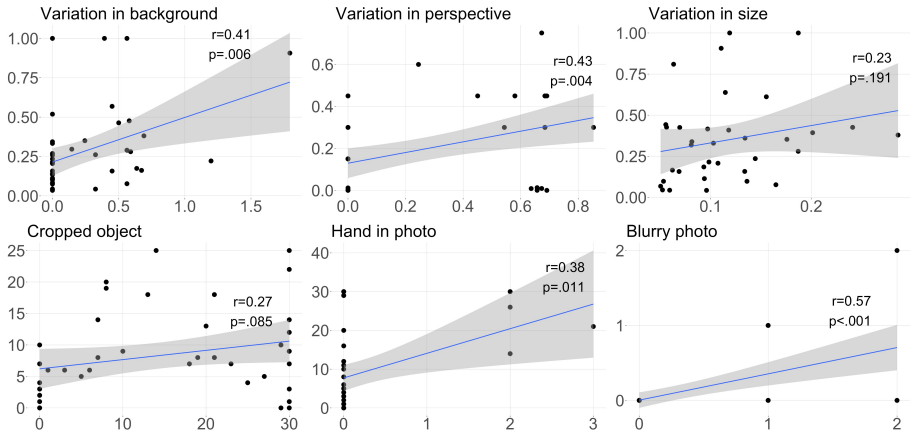
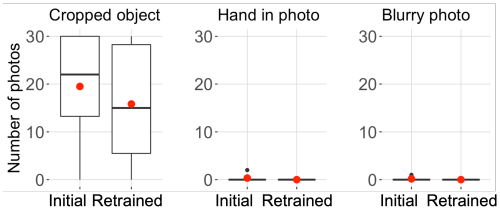
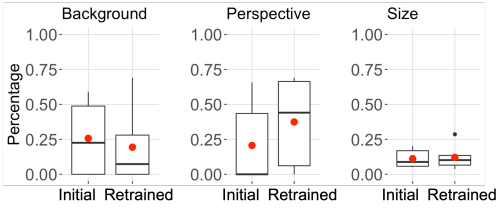


Fig. 6. Scatter plots indicating correlations between manual annotations (x-axis) and estimations (y-axis) for each descriptor.



(a) The aggregate of photo-level descriptors.



(b) Set-level descriptors.

Fig. 7. Contrasting descriptor values in initial attempts to retraining attempts for P1, P3, P5, P8, and P10. Red dots indicate means.

5.2 Changes in Annotated Descriptors For Participants who Choose to Retrain

Five participants (P1, P3, P5, P8, P10) decided to retrain with a new set of photos for an object after reviewing their initial training sets; one of them (P3) trained the same object three times, each time with a new set of photos. A participant (P10) retrained with new sets of photos for two of the three objects. No participant retrained all three objects.

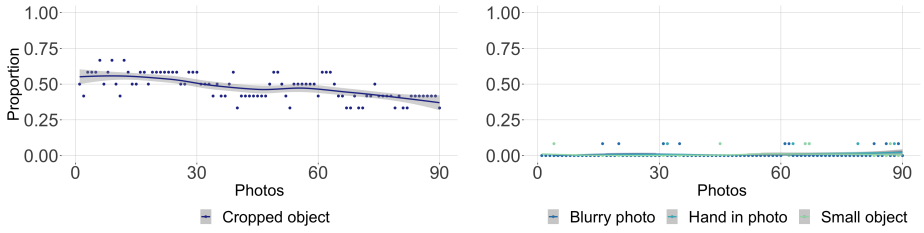
As shown in Figure 7, we contrast the estimated descriptors for initial attempts to those during retraining attempts. When the participants decided to retrain, their new training sets had fewer photos with cropped objects, no hands included, almost no blurred photos, and higher variation in perspective and size on average compared to their initial photos. This is a promising trend providing some evidence on participants' attempt to respond and adhere to the descriptors though it may have come at the cost of lower variation for background.

Specifically, the average numbers of photos with cropped objects and users' hands were fewer at 15.83 (SD = 13.41) and 0.00 (SD = 0.00) in their new training photos versus the initial at 19.50 (SD = 12.52) and 0.33 (SD = 0.81), respectively. The number of blurry photos was 0.00 (SD = 0.40) and 0.17 (SD = 0.41) in retrained and initial, respectively. The system did not detect any photos with a too-small object in either set. As for variation, mean variation in perspective and size in retrained were 0.37 (SD = 0.33) and 0.12 (SD = 0.09), respectively, which is higher compared to those in the initial sets at 0.20 (SD = 0.32) and 0.11 (SD = 0.07). However, this trend was reversed for variation in background. This descriptor was on average lower in retrained at 0.19 (SD = 0.28) compared to the initial at 0.26 (SD = 0.28).

5.3 Changes in Annotated Attributes for All Participants Over Time

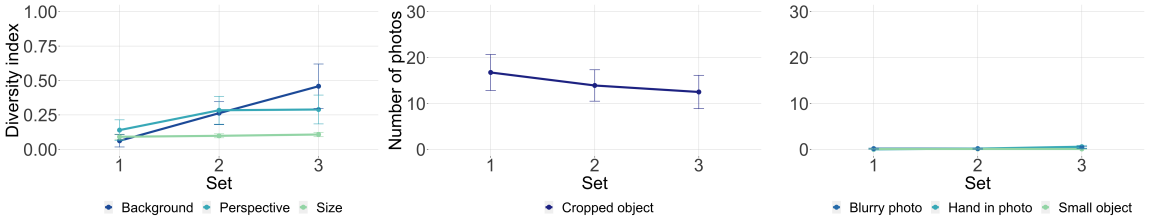
Many participants chose not to retrain. Perhaps the interactive nature of the descriptors created opportunities for early reflection and experimentation; not just at the end of training. To explore this, we measured trends over time at different levels of granularity; for this analysis, we use the manually annotated attributes, which serve as the ground truth, rather than the estimated descriptors.

5.3.1 Fine-grained Changes Across 90 Training Photos. With photo-level descriptors participants' could gauge potential image quality issues right away; MYCam indicates them immediately after a photo is taken. As shown in Figure 8a, we observe a dropping trend in the number of images where the object was cropped as participants progressed in the study. This is promising for a descriptor that merely provides binary feedback (*i.e.*, whether the object is cropped or not) instead of directional guidance on how to move a camera to fully capture an object (*e.g.*, Lee *et al.* [41]). The proportion of photos with cropped objects was around 0.56 at the beginning (1st photo in 1st training), decreasing to 0.37 by the last photo (30th photo in 3rd training). Whereas the proportion of training examples with participants' hands included, objects too small, or blurry photos were nearly zero throughout the study (Figure 8b).



(a) Cropped object. (b) Blurry photo, hand in photo, too small object.

Fig. 8. The average values of annotated photo-level attributes for individual photos among 12 participants. The charts include photos of the first three training sets (1-30: first set, 31-60: second set, 61-90: third set). The lines are fitted to dots using LOWESS smoothing.



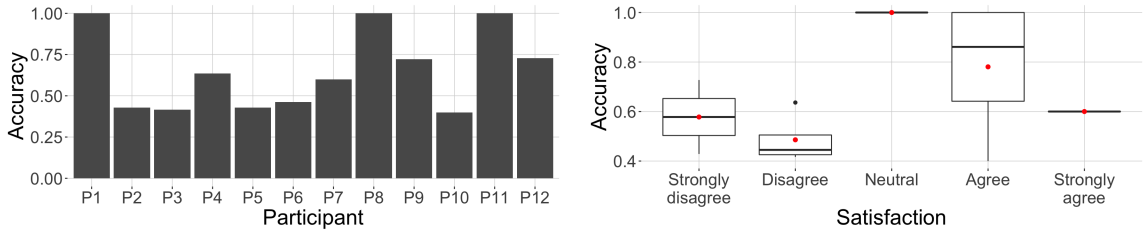
(a) Set-level descriptors. (b) Cropped object. (c) Blurry photo, hand in photo, small object.

Fig. 9. The average annotated values of set-level attributes and the annotated number of photos with photo-level attributes for all 12 participants across three training sets (a training set per object).

5.3.2 Coarse-grained Changes Across 3 Training Sets. With photo-level descriptors and set-level aggregates participants' could gauge potential issues related to their teaching strategies or image quality at the end of each training attempt; MYCam shows them immediately after 30 photos are taken. Participants may or may not choose to go back and retrain. But they may also choose to reflect when training the next object, especially since our object stimuli were engineered to be very similar. As shown in Figure 9, participants increased the variation among their training examples and reduced the number of photos with cropped objects. A one-way repeated-measure ANOVA indicate a significant effect of order of sets on variation in background ($F(2, 22) = 4.59, p = 0.022$, partial $\eta^2 = 0.18$) and in perspective ($F(2, 22) = 3.61, p = 0.044$, partial $\eta^2 = 0.05$). We did not observe a statistically significant effect of the other attributes. However, we do observe a tendency for an increase in the number of photos that were blurry or where the participant's hand was included. Perhaps these descriptors were not deemed as that problematic or they were ranked lower in priority as teaching strategies evolved. Participants' feedback below can shed a bit more light on these observations.

5.4 Performance of Participants' Object Recognition Models

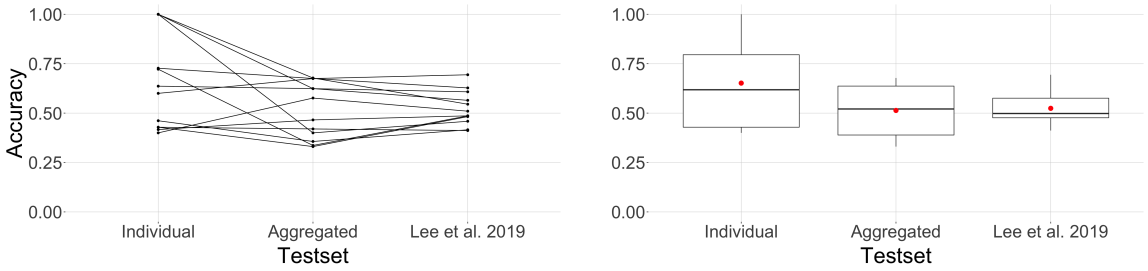
After finalizing their training for all objects, participants were called to test the performance of their models; we explicitly did not allow for intermediate train-test iterations in an attempt to limit interference from that type of experimentation in the observed behaviors. For the purpose of our analysis, we report model performance not only on participants' final training sets but also dive deeper and look at their photos chosen to test their models and how well their model generalizes *e.g.*, if tested with photos taken by others.



(a) Object recognition accuracy on one's own testing images.

(b) Accuracy against participant satisfaction with performance.

Fig. 10. When testing their models, participants' experiences varied (a), which seems to be reflected in their satisfaction scores (b).



(a) Accuracy per participant.

(b) Summary statistics of accuracy.

Fig. 11. Model accuracy when tested on individual test images, aggregated test images from all 12 blind participants in this remote study, and aggregated test images from all 9 blind participants in a prior in-lab study [41].

5.4.1 Model Performance with Testing Images from Self. We found that participants used a very small number of photos ($M = 3.7$, $SD = 3.2$) to check if their models were working properly. Some (4 out of 12) included photos where the object was more than half cropped. Others (4 out of 12) captured multiple objects in the frame. Some of these observations could be perhaps explained by our study setup (*e.g.*, participants were done with taking photos for the day or objects were in close proximity due to study setup). However, prior work in teachable object recognizers employing different study designs also indicates that model testing and evaluation can be challenging for end users [17, 27]. These challenges are critical as perceived and actual performance may be different when the models are actually used after testing.

Overall, we find that when testing on one's own data the average accuracy (*i.e.*, the number of correct predictions divided by total test images) of the models was 0.65 ($SD = 0.24$) with a breakdown across participants shown in Figure 10a. These results may seem surprisingly low for a 3-way classification task. However, beyond being a fine-grained classification, the task can be particularly challenging with objects of deformable shapes, same-size, reflective-surface, and similar colors that can be hard to distinguish. Among the high-performing models are those of P1 and P8 who choose to iterate on their training (they tested the models with 3 and 7 photos, respectively); though the same is not to be said for the models of P3, P5, and P10 who also iterated on their training (they tested the models with 12, 28, and 10 photos, respectively). When juxtaposing model performance with participants' subjective responses on the satisfaction of their models (Figure 10b), we find that those whose models did not perform well disagree with this statement and those whose models perform better agree. This alignment, however, did not hold for those on the edges (*strongly disagree* and *strongly agree*). Participants' feedback in the next section, provides a potential explanation.

5.4.2 Model Performance with Testing Images from Others. One of the promises of a teachable object recognizer is that it works well for each individual since the training and test sets are collected by the same person and it is highly likely that they are going to exhibit similar patterns [34, 68]. This was also the case in our study. As shown in Figure 11, for 9 out of 12 participants, the accuracy of the model was higher when tested with an individual participant's test set than an aggregated test set from all participants in our study and photos from another study with blind participants [41] on the same objects. The accuracy of the model with individual test sets was 0.65 (SD = 0.24). The accuracy was lower at 0.51 (SD = 0.14) and 0.52 (SD = 0.09) when pooling test sets across all participants in the current study and testing photos from a prior study [41], where nine blind participants trained and tested a teachable object recognizer, respectively. However, we observed that the iteration can make the models generalize better. Among the five participants who did retraining, four and three participants had higher accuracy after retraining when their models were tested with the aggregated test set and the set from the prior study [41], respectively.

5.5 Subjective Feedback from Participants

5.5.1 Overall Experience. To provide more context on participants' feedback for the descriptors, we illustrate in Figure 12 their responses related to the MYCam testbed. Overall, participants agreed that they could train the object recognition model effectively with MYCam and disagreed on training being difficult, though they were divided on whether it could be done quickly. This is promising. Specifically, ten participants agreed or strongly agreed that they could train their models effectively with some pointing to the need for onboarding. P1 and P10, for example, who are not familiar at all and slightly familiar with machine learning said *"after a while, I learned that I could train it"* and *"It's pretty easy. You have to teach me though. But if you teach me then it's pretty easy to follow instructions and finish the process."* respectively. On the other hand, P11 and P12 were neutral. P11 mentioned that taking 30 photos is time-consuming, saying *"I don't really feel like I was all that effective because it takes a while to train for each one."* The errors in descriptors affected the reliability of the app, making a participant think the training process was less effective even though the two models work independently of each other. P12 said *"I don't think that the app is correct, especially when I know, for example, that my hand was not in the photo...I don't have a lot of confidence in the app's accuracy."*

When asked whether they could train the app quickly, five participants agreed, four disagreed, and three were neutral. Seven participants indicated that taking 30 photos is tedious. For example, P10 said, *"The process is pretty straightforward. But I have to spend, like, quite long time to train the three objects."* When asked about the difficulty of the training task, all but one who remained neutral, disagreed or strongly disagreed that the task was difficult. P11 who was neutral, found it not difficult but tedious. Surprisingly, this sentiment of the task being tedious was not presented in the initial study with teachable object recognizers [34] even though the number of training photos was identical. We suspect this difference reflects more on our implementation of the descriptors in the MYCam testbed rather than the process of training itself. In MYCam users could not opt in or out of the photo-level descriptors during training leading to higher training times; specifically, the time for taking photos for training an object was doubled from 65 seconds (SD = 35.2) reported in that first study [34] to 143.8 seconds on average (SD = 72.4) in our study.

5.5.2 Descriptors. As shown in Figure 13, all but one participant (P1, who was neutral) agreed or strongly agreed that the **descriptors were easy to understand**. P6 said *"I understood what it was telling me. I didn't have questions about what I was supposed to do."* Participants' responses indicate user reflection based on descriptor changes across multiple attempts, strengthening some of our observations in the previous sections. P2 elaborated *"It gives you directions. The explanation (descriptors) afterwards, in the analysis, told me that my photographs were not always good. So I have to*

learn to take better photographs." Some participants found it difficult to understand the absolute values of provided in the descriptors and were wondering whether they should have a specific value as a goal. For example, the values of descriptors were somewhat ambiguous to P1 who said, *"I guess just knowing exactly what they're referring to what numbers are really preferable."* P4 also mentioned the challenge in understanding the values of descriptors, but then mentioned that over repeated data collection during the study, he figured out their purpose. P4 said *"I wasn't aware of any of those fields when we did the first object [...] For the second and third objects, I could take a little bit more variation in the photos or to better train the application."* This is interesting feedback as the descriptors are there merely to provide access to what one could infer via a visual inspection not per se dictate optimal characteristics for the training set. The difference of course is that when a sighted person glances over their training photos they may or may not make an inference on potentially problematic photos or lack of variation (see Hong *et al.* [27]), but a blind person always hears the descriptors. This explicit presence of the descriptors calls for the need for more context. While "ideal values" are use-case depended, during onboarding users could perhaps be provided with some rationale or examples.

Ten participants agreed or strongly agreed that the **descriptors were useful**. P10 (who was neutral on this) and P11 thought descriptors helped them understand how to collect training examples for the object recognizer. P10 said, *"(I agree) because I know the quality of the photos, the different aspects of the photos that I take."* P11 said, *"It helped me understand what the camera needed in order to recognize the objects."* Participants also mentioned that descriptors were useful to identify problems in their training sets. P10 elaborated *"you have to get feedback or you're not going to improve [...] it helps you to understand what you're doing wrong."* P2 had a similar idea: *"the explanation (descriptors) afterward, in the analysis, told me that my photographs were not always good, so I have to learn to take better photographs."* P12 thought they were not useful because they were error-prone. P12 said, *"I don't think that the app is correct, especially when I know, for example, that my hand was not in the photo, or that the object is not cropped because the previous objects were cropped."* This feedback highlights the need for improving the estimation of descriptors in future work.

Participant feedback suggests that it would have been helpful to include more explicit guidance on how to improve the training photos. For example, P7 suggested similar feedback to Lee *et al.* [41] and Ahmetovic *et al.* [3] along the

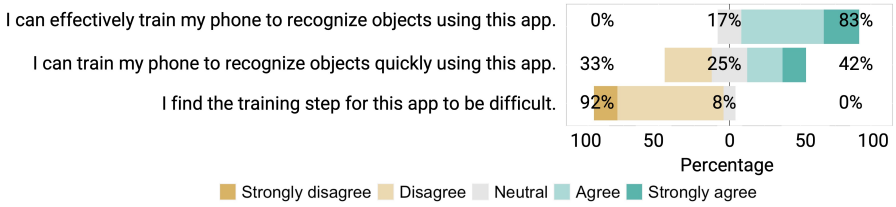


Fig. 12. Participants' feedback on training with the MYCam testbed.

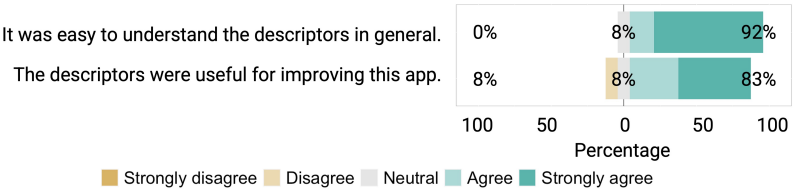


Fig. 13. Participants' feedback on the descriptors.

descriptors, elaborating *"Cropped, it did not help me know what to do differently. If it said, maybe move up, move down and move camera left, move the camera, right. That would have been more useful."* Our current implementation of this photo-level descriptor actually can be re-purposed to provide such feedback. More so, P6 mentioned that the interface for replacing problematic photos in a training set would improve the app. He said *"I would assume the training process can self-evaluate itself and it should sum that up for me and tell me what photos I should replace. [...] you need to replace those bad pictures unless you don't need them for the training."* This is an intriguing approach, one that we aim to explore.

5.5.3 Model Performance. When we asked participants if they were satisfied with the performance of the object recognizer, opinions were divided; five participants agreed or strongly agreed, six participants disagreed or strongly disagreed, and one participant was neutral, as shown in Figure 14. When accounting for the performance of their model in their subjective responses (Figure 10b), we observe that participants were not satisfied if the accuracy was lower than 0.6. However, it did not all come down to model performance. Open-ended feedback indicates that satisfaction is also related to effort. P11 remained neutral even though she did not observe any recognition errors (her model had the highest accuracy) but attributed this to the training task being tedious (see Section 5.5.1). P11 said, *"Because it took so much work to get that small amount of performance."*

P7 and P10 agreed that they were satisfied with the performance though their model accuracy was only 0.6 and 0.4, respectively. As legally blind P7, expressed that the performance is good enough to supplement his vision. P10 believed that she just did not train the app properly. She said, *"I think it recognized objects, but if you don't train it properly, then it's not going to recognize anything [...] the Fritos bag was the one that didn't work out, but that was probably my fault."*

While the majority (9 out of 12) of participants observed recognition errors during testing, many could not explain why. Six participants were neutral or disagreed with the statement that they have a good sense of why the recognition errors occurred. Their responses were simply *"I have no idea."* or *"I don't know."* Though P7 and P10 strongly agreed and agreed, respectively, their rationale was vague. P10 said *"I think it was my fault. I think it was my training. Other than that, I don't know."* P9 strongly agreed and contributed the recognition errors to imperfect descriptors in training, elaborating *"The reason is because I was teaching it, and I wasn't 100% sure that it was 100% accurate. It makes sense that while I was teaching it, I was a little bit off, so its recognition was a little bit off. It kept telling me that the hand was in the photos."* We believe that these observations motivate the need for accessible computer vision explanations.

6 DISCUSSIONS

Our user study, exploratory in nature, shows both promising results and future research directions for supporting blind users' interactions with teachable machines. In this section, we first reflect on lessons learned, while discussing

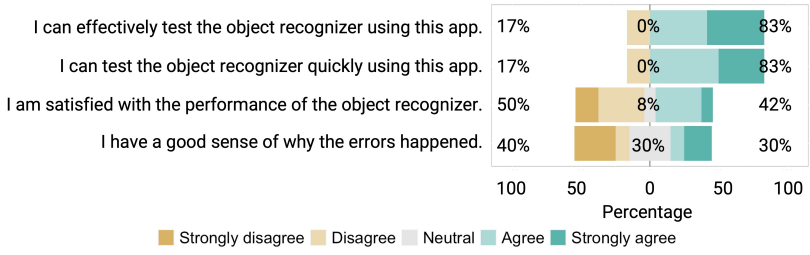


Fig. 14. Participants' feedback on the performance of their object recognition models.

implications for designing descriptors to access one's training data in teachable object recognizers and broader teachable applications either assistive or educational. We then discuss limitations in our study that may affect the generalizability of our findings as well as future work for better estimating such descriptors and exploring their potential for explainability.

6.1 Implications

Our study provides evidence that descriptors derived from visual attributes used to code training photos in teachable object recognizers, can provide blind users with a means to inspect their data, iterate, and improve their training examples. Challenges often involve onboarding, time needed for training, as well as descriptor accuracy and interpretation.

Insights from this work are complementary to prior studies exploring the feasibility of training [34, 47, 68] and camera aiming [3, 42, 72] in teachable object recognizers for the blind. More so, the underlying methods for extracting meaningful descriptors, *i.e.*, instance- and set-level characteristics that can be coded by quickly inspecting the training data and that point to noise and variation, respectively, can be adopted for other teachable applications. This is especially critical for those assistive applications where training typically requires similar skills to those the technology aims to fulfill. For example, teachable sound detectors for Deaf/deaf and hard of hearing people [9, 20] could benefit from visualizations of the sound examples in a way that allows users to quickly inspect potential noise in a training example and variation across the training set (*e.g.*, better start and end of a recording, multiple sound sources, variation, and other characteristics that hearing users could leverage for experimentation just by listening to the audio). Indeed, Goodman *et al.* [20] observed that Deaf/deaf users collected similar-sounding examples during training and thus, could benefit from an interface that visualizes features of their training sets; data descriptors could fill that need.

Accessing one's training data is also critical for making informal learning activities that typically employ teachable machines with children more inclusive. Learning objectives for AI education in K12 (*e.g.* [73]) highlight the use of interactive systems for exposing children to AI prior to using those that leverage block-based programming [74]. Dwivedi *et al.* [17] suggest that future teachable interfaces for such activities benefit from classification tasks that allow children to quickly inspect the data and uncover patterns. Thus, it is not a surprise to see many learning activities for exposing children to AI often leverage teachable image classification applications [14, 17, 22, 39, 75]. However, in these initial explorations, none of these applications are inclusive of blind children. Our data descriptors could help increase their accessibility *e.g.*, by leveraging our shared code for MYCam and the descriptors. Further, we see how researchers working in teachable object recognizers and broader contexts, could benefit from the following insights:

Balancing descriptors with demand on time and cognitive load. MYCam's set-level descriptors are given at the end of training but image-level descriptors are played every time the user takes a training photo (they can also be accessed when reviewing at a later time). Participants' feedback indicates that, although the image-level descriptors are informative, they add to the training time and cognitive load. Indeed, if we were to compare our study with the times reported in Kacorri *et al.* [34] training with descriptors (4.8 seconds per photo on average) took more than double the time without them (2.2 seconds on average), respectively. Still, this was much less when compared to another study, where blind users took photos with real-time camera aiming guidance (*i.e.*, audio and haptic feedback for camera aiming); there, they spent on average 10.3 seconds per photo [41] but did not reflect much on the time needed to train. The difference between the two is: MYCam feedback when taking photos is passive and requires listening to a list of descriptors and optimizing simultaneously multiple variables whereas the feedback in Lee *et al.* [41] is interactive and requires listening to an audio cue or sensing vibration and optimizing a single variable (including the object in the frame). This challenge of maximizing

information while minimizing cognitive load is not new and calls for better interactivity with the descriptors via opt in/out mechanisms (*e.g.*, play descriptors via press and hold), verbosity controls, or audio haptic feedback. For example, P10 expressed that while descriptors provide hints to problems, they do not directly instruct users on how to solve them. For example, when an object is cropped in a photo, participants did not get feedback on in which direction the camera should move even though this information can be made available from the current implementation. We expect that combining the descriptors with camera guidance (*e.g.*, [3, 41]) could be helpful.

Balancing descriptors with instructions. There is a rich literature on the value of tutorials, instructions, and in-context interactive assistance for supporting users with technology; a comprehensive review for blind users and smartphone devices can be found in Rodrigues *et al.* [59]. Some prior studies with blind users have shown that real-time descriptions can lead to better accuracy and confidence compared to instructions at the start of a task (*e.g.*, [19]). While we did not compare the two, participants' feedback indicate that data descriptors would be complementary and not a substitute for tutorials and instructions. In addition to the real-time feedback, participants call for support in navigating the app and interpreting descriptor values. In our study, the experimenter provided some of this information. For example, for the set-level descriptors the experimenter said: "you can check how much variation your photos have. For example, a 10% variation in the background means that most of your photos have similar backgrounds." However, participants mentioned that they could better understand the absolute values of descriptors after experimenting. We suspect that the level of understanding for these values would affect both the quality of the training sets as well as how reliable the system is perceived by the users.

Editing a training set based on descriptors. The current design of MYCam focuses on informing the users of the attributes of their training sets rather than instructing how to spot potential issues in their training sets or making the data collection process efficient. Participants had to diagnose problems for themselves based on descriptors and replace the entire training set with new photos if they wanted to fix something. Participants suggested adding functions to edit (*e.g.*, delete) at a photo level *e.g.*, right after taking a photo that is deemed noisy or while reviewing the training set at the end to make the iterations more efficient. For example, P8 suggested having an interface that filters out bad images based on descriptors or enables users to replace them instead of starting from scratch. This opens up interesting venues for approaches such as active learning and data valuation.

6.2 Limitations and Future Work

There are many limitations that could impact the generalizability of our findings. Our observations come from a small sample even though $N = 12$ is most common in human-computer interaction studies [12]. Our study is remote, yet participants are recruited from a relatively small area in the US in proximity to the authors' institution. The study is conducted in participants' homes, yet it shares more characteristics with a controlled in-lab study rather than a real-world deployment: object stimuli were predefined and small in number, the duration of the study was relatively short, MYCam was deployed on one of our devices, participants were being observed and had real-time support from the experimenter, and they were somewhat confined in terms of space.

Specifically, participants were asked to wear smart glasses and communicate with the experimenter through a laptop computer in front of them. Though these devices were necessary for communication and data analysis, they limit participants' behavior *e.g.*, in walking around with the phone and taking photos in different locations and illuminations. For example, when participants wanted to vary the backgrounds in photos, they took pictures with different parts of a table. However, if they could move around outside the user study setup, perhaps they would choose completely different locations for background variation. More so, using MYCam on one of our iPhone 8 devices instead of their own mobile

devices could have affected our observations. All but one participant owned an iPhone; most participants were familiar with iOS apps. However, the difference between their personal phones and our device (*e.g.*, in terms of size and camera location) could have affected the quality of photos and overall perception of the descriptors. We expect that the use of MYCam in a real-world scenario would have resulted in a richer set of contexts in users' photos (typically a table in our study). Though we limited the objects to three snacks with similar textures and weights, blind people may choose to train on personal objects that may not be products in the market with a larger number of object instances. As the performance of an object recognizer depends on the number of classes and visual difference between the objects, these differences could have affected the performance of a personalized model and blind users' experiences with it.

Our experimental setup was in part restricted by our implementation of the descriptors, which is meant to serve as a proof of concept and is somewhat tied to a predefined set of object stimuli (*e.g.*, for the ARKit to work and for establishing different thresholds). Although the estimated descriptors had a positive correlation with the manually annotated attributes and enabled participants to inspect their training sets, they were error-prone. When some of the participants noticed the errors in descriptors, they deemed them as well as the object recognition model unreliable. This suggests that it is imperative to further advance approaches related to descriptor estimation for a better user experience.

Due to the lack of datasets for benchmarking our descriptor-based approach, we had to manually create our own dataset for comparison. As in other AI-based systems evaluation, having benchmark datasets is useful to assess systems for generating descriptors in a more widely accepted way. One potential step in this direction would be to invite blind data contributors, who can inspect their personal training data and agree to data sharing, to contribute to such benchmark datasets employing approaches similar to Theodorou *et al.* [72].

Last, in this study, images were used for the purpose of training. This approach can provide more control for the blind users over their training sets regarding both incorporating variations and mitigating privacy risk concerns [5] as it would be less likely for a blind user to capture unnecessary information in an image. For example, blind users usually use their hands as a reference to center the object in the camera frame, but they are often not willing to include their hands in the final photo to preserve user anonymity [41]. Also, in the one study where videos were used, blind users had to be trained to follow some instructions and filming techniques [72]. On the other hand, video increases the number of collected images since it is a collection of frames. Also, the use of video increases the chance of the object being in the frame at some point [72]. Perhaps a way to get the best of the two worlds could be live photos as they are easy to capture (like photos), and they include multiple frames over one to three seconds [53].

7 CONCLUSION

In this work, we examined the challenge of accessing one's training examples in teachable object recognizers, where visual inspection of training photos is not accessible to blind users with the ultimate goal of making machine teaching more inclusive. To this end, we engineered real-time descriptors that indicate to the blind user whether the photo they just took is blurry, if their hand is in it, if the object is cropped, and whether their photos overall vary in object background, distance, and perspective; all factors that can affect model performance. We built MYCam, an accessible and open-source teachable object recognizer iOS app with descriptors. We shared our findings, observations, and lessons learned from a remote study with 12 blind participants who trained MYCam in their homes to recognize three distinct but visually similar objects.

Our results showed that participants who choose to iterate their training for an object, were able to provide fewer photos where the object was cropped, included no hand in their photos, and had slightly less blurry photos that overall had more variation in terms of object perspective and size but less in terms of background. Overall, participants

increased the variation among their training examples and reduced the number of photos with cropped objects as they moved in training from one object to the next. Some of these changes are reflected in their model performance that somewhat relate to their satisfaction scores. However, errors in descriptor estimates seem to affect overall participants' perception and trust of model performance. Participants' responses indicate that even though it was difficult to gauge the meaning of absolute values for some of the descriptors (*e.g.*, variation), they could infer it based on relative changes. However, many found the training being tedious, opening discussions around the need for balance between information, time, and cognitive load. These results, taken together, indicate that our novel data descriptors, realized in MYCam, hold potential for facilitating quick inspection of training photos among blind individuals. Going forward, we are excited to continue our endeavors towards building more inclusive participatory machine learning experiences both for blind youth and adults.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their thoughtful comments on an earlier draft of this paper. This work is supported by NSF (1816380). Kyungjun Lee is supported by NIDILRR (90REGE0008).

REFERENCES

[1] Ali Abdolrahmani, William Easley, Michele Williams, Stacy Branham, and Amy Hurst. 2017. Embracing Errors: Examining How Context of Use Impacts Blind Individuals' Acceptance of Navigation Aid Errors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4158–4169. https://doi.org/10.1145/3025453.3025528

[2] Dragan Ahmetovic, Cristian Bernareggi, Andrea Gerino, and Sergio Mascetti. 2014. ZebraRecognizer: Efficient and Precise Localization of Pedestrian Crossings. In *2014 22nd International Conference on Pattern Recognition*. 2566–2571. https://doi.org/10.1109/ICPR.2014.443

[3] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. *ReCog: Supporting Blind People in Recognizing Personal Objects*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376143

[4] Aira. 2017. *Your Life, Your Schedule, Right Now.* https://aira.io

[5] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. 2020. "I am uncomfortable sharing what I can't see": Privacy Concerns of the Visually Impaired with Camera Based Assistive Applications. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 1929–1948. https://www.usenix.org/conference/usenixsecurity20/presentation/akter

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[7] Matthias Bauer, Mateo Rojas-Carulla, Jakub Bartłomiej Świątkowski, Bernhard Schölkopf, and Richard E Turner. 2017. Discriminative k-shot learning using probabilistic models. *arXiv preprint arXiv:1706.00326* (2017). https://doi.org/10.48550/ARXIV.1706.00326

[8] BeMyEyes. 2016. *Lend you eyes to the blind.* http://www.bemyeyes.org/

[9] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. 2016. A Personalizable Mobile Sound Detector App Design for Deaf and Hard-of-Hearing Users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 3–13. https://doi.org/10.1145/2982142.2982171

[10] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 16–31. https://doi.org/10.1145/3308561.3353774

[11] John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, and Richard Turner. 2020. TaskNorm: Rethinking Batch Normalization for Meta-Learning. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 1153–1164. https://proceedings.mlr.press/v119/bronskill20a.html

[12] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[13] Kathleen Campbell, Kimberly LH Carpenter, Jordan Hashemi, Steven Espinosa, Samuel Marsan, Jana Schaich Borg, Zhuoqing Chang, Qiang Qiu, Saritha Vermeer, Elizabeth Adler, Mariano Tepper, Helen L Egger, Jeffery P Baker, Guillermo Sapiro, and Geraldine Dawson. 2019. Computer vision analysis captures atypical attention in toddlers with autism. *Autism* 23, 3 (2019), 619–628. https://doi.org/10.1177/1362361318766247 arXiv:https://doi.org/10.1177/1362361318766247 PMID: 29595333.

[14] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382839

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[16] Voice Dream. 2022. *Scanner – Voice Dream.* https://www.voicedream.com/scanner/

[17] Utkarsh Dwivedi, Jaina Gandhi, Raj Parikh, Merijke Coenraad, Elizabeth Bonsignore, and Hernisa Kacorri. 2021. Exploring Machine Teaching with Children. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 1–11. https://doi.org/10.1109/VL/HCC51201.2021.9576171

[18] Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. 2014. Headlock: A Wearable Navigation Aid That Helps Blind Cane Users Traverse Large Open Spaces. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers amp; Accessibility* (Rochester, New York, USA) *(ASSETS '14)*. Association for Computing Machinery, New York, NY, USA, 19–26. https://doi.org/10.1145/2661334.2661453

[19] Nicholas A. Giudice, Benjamin A. Guenther, Toni M. Kaplan, Shane M. Anderson, Robert J. Knuesel, and Joseph F. Cioffi. 2020. Use of an Indoor Navigation System by Sighted and Blind Travelers: Performance Similarities across Visual Status and Age. *ACM Trans. Access. Comput.* 13, 3, Article 11 (aug 2020), 27 pages. https://doi.org/10.1145/3407191

[20] Steven M. Goodman, Ping Liu, Dhruv Jain, Emma J. McDonnell, Jon E. Froehlich, and Leah Findlater. 2021. Toward User-Driven Sound Recognizer Personalization with People Who Are d/Deaf or Hard of Hearing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 63 (jun 2021), 23 pages. https://doi.org/10.1145/3463501

[21] Google. 2022. *Lookout - Assisted vision - Apps on Google Play.* https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US&gl=US

[22] Google. 2022. Teachable Machine. https://teachablemachine.withgoogle.com/.

[23] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M. Kitani, and Chieko Asakawa. 2019. CaBot: Designing and Evaluating an Autonomous Navigation Robot for Blind People. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 68–82. https://doi.org/10.1145/3308561.3353771

[24] Anhong Guo, Xiang 'Anthony' Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P. Bigham. 2016. VizLens: A Robust and Interactive Screen Reader for Interfaces in the Real World. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 651–664. https://doi.org/10.1145/2984511.2984518

[25] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300645

[26] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2019. Exploring Machine Teaching for Object Recognition with the Crowd. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312873

[27] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. *Crowdsourcing the Perception of Machine Teaching.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376428

[28] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers* 12, 4 (2000), 409–426. https://doi.org/10.1016/S0953-5438(99)00006-5

[29] Virtual Collaboration Research Inc. 2022. *Super Lidar - Lidar for Blind.* https://apps.apple.com/us/app/super-lidar-lidar-for-blind/id1543706309

[30] Hairong Jiang, Ting Zhang, Juan P Wachs, and Bradley S Duerstock. 2016. Enhanced control of a wheelchair-mounted robotic manipulator using 3-D vision and multimodal interaction. *Computer Vision and Image Understanding* 149 (2016), 21–31. https://doi.org/10.1016/j.cviu.2016.03.015

[31] Hernisa Kacorri. 2017. Teachable Machines for Accessibility. *SIGACCESS Access. Comput.* 119 (nov 2017), 10–18. https://doi.org/10.1145/3167902.3167904

[32] Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka Jha, and Riya Chanduka. 2020. IncluSet: A Data Surfacing Repository for Accessibility Datasets. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 72, 4 pages. https://doi.org/10.1145/3373625.3418026

[33] Hernisa Kacorri, Utkarsh Dwivedi, and Rie Kamikubo. 2020. Data Sharing in Wellness, Accessibility, and Aging. *NeurIPS 2020 Workshop on Dataset Curation and Security* (2020).

[34] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5839–5849. https://doi.org/10.1145/3025453.3025899

[35] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300282

[36] Taha Khan, Dag Nyholm, Jerker Westin, and Mark Dougherty. 2014. A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artificial intelligence in medicine* 60, 1 (2014), 27–40. https://doi.org/10.1016/j.artmed.2013.11.004

[37] Ron Kimmel and Alfred M Bruckstein. 2003. Regularized Laplacian zero crossings as optimal edge integrators. *International Journal of Computer Vision* 53, 3 (2003), 225–243. https://doi.org/10.1023/A:1023030907417

[38] Masaki Kuribayashi, Seita Kayukawa, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2021. LineChaser: A Smartphone-Based Navigation System for Blind People to Stand in Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 33, 13 pages. https://doi.org/10.1145/3411764.3445451

[39] Google Creative Lab. 2017. Teachable Machine. https://teachablemachine.withgoogle.com/v1/.

[40] Kyungjun Lee, Jonggi Hong, Ebrima Jarjue, Ernest Essuah Mensah, and Hernisa Kacorri. 2022. From the Lab to People's Home: Lessons from Accessing Blind Participants' Interactions via Smart Glasses in Remote Studies. In *Proceedings of the 19th International Web for All Conference* (Lyon, France) *(W4A '22)*. Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. https://doi.org/10.1145/3493612.3520448

[41] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting Blind Photography in the Context of Teachable Object Recognizers. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 83–95. https://doi.org/10.1145/3308561.3353799

[42] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300566

[43] Kyungjun Lee, Daisuke Sato, Saki Asakawa, Hernisa Kacorri, and Chieko Asakawa. 2020. *Pedestrian Detection with Wearable Cameras for the Blind: A Two-Way Perspective*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376398

[44] James R Lewis. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78. https://doi.org/10.1080/10447319509526110

[45] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[46] Cristina Manresa-Yee, Javier Varona, Francisco J Perales, and Iosune Salinas. 2014. Design recommendations for camera-based head-controlled interfaces that replace the mouse for motion-impaired users. *Universal access in the information society* 13, 4 (2014), 471–482. https://doi.org/10.1007/s10209-013-0326-z

[47] Daniela Massiceti, Lida Theodorou, Luisa Zintgraf, Matthew Tobias Harris, Simone Stumpf, Cecily Morrison, Edward Cutrell, and Katja Hofmann. 2021. ORBIT: A real-world few-shot dataset for teachable object recognition collected from people who are blind or low vision. https://doi.org/10.25383/CITY.14294597

[48] Mediate. 2022. *Supersense - AI for Blind / Scan text, money and objects.* https://www.supersense.app/

[49] Carrie Morales. 2019. *What's Better for the Blind and Low Vision? Android or iPhone?* https://liveaccessible.com/2019/03/03/whats-better-for-the-blind-and-low-vision-android-or-iphone/

[50] John Morris and James Mueller. 2014. Blind and deaf consumer preferences for android and iOS smartphones. In *Inclusive designing*. Springer, 69–79. https://doi.org/10.1007/978-3-319-05095-9_7

[51] Meredith Ringel Morris. 2020. AI and Accessibility. *Commun. ACM* 63, 6 (2020), 35–37. https://doi.org/10.1145/3356727

[52] Donald A Norman. 1994. How might people interact with agents. *Commun. ACM* 37, 7 (1994), 68–71. https://doi.org/10.1145/176789.176796

[53] Lauren Olson, Chandra Kambhamettu, and Kathleen McCoy. 2021. *Towards Using Live Photos to Mitigate Image Quality Issues In VQA Photography*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3441852.3476541

[54] Thomas J. Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience* 5, 4 (2004), 291–303. https://doi.org/10.1038/nrn1364

[55] Rupal Patel and Deb Roy. 1998. Teachable interfaces for individuals with dysarthric speech and severe physical disabilities. In *Proceedings of the AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*. Citeseer, 40–47.

[56] Rubens Lacerda Queiroz, Fábio Ferrentini Sampaio, Cabral Lima, and Priscila Machado Vieira Lima. 2020. AI from concrete to abstract: demystifying artificial intelligence to the general public. https://doi.org/10.1007/s00146-021-01151-x arXiv:2006.04013 [cs.CY]

[57] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. https://doi.org/10.48550/ARXIV.1804.02767

[58] Alejandro Reyes-Amaro, Yanet Fadraga-González, Oscar Luis Vera-Pérez, Elizabeth Domínguez-Campillo, Jenny Nodarse-Ravelo, Alejandro Mesejo-Chiong, Biel Moyà-Alcover, and Antoni Jaume-i Capó. 2012. Rehabilitation of patients with motor disabilities using computer vision based techniques. *Journal of accessibility and design for all* 2, 1 (2012), 62–70. https://doi.org/10.17411/jacces.v2i1.87

[59] André Rodrigues, André Santos, Kyle Montague, Hugo Nicolau, and Tiago Guerreiro. 2019. Understanding the Authoring and Playthrough of Nonvisual Smartphone Tutorials. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 42–62.

[60] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 222–235. https://doi.org/10.1145/3308561.3353776

[61] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. In *HCOMP*.

[62] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *CoRR* abs/1708.08296 (2017). arXiv:1708.08296 http://arxiv.org/abs/1708.08296

[63] SeeingAI. 2017. *An app for visually impaired people that narrates the world around you.* https://www.microsoft.com/en-us/seeing-ai

[64] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* 5, 1 (2001), 3–55.

[65] Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. https://doi.org/10.48550/ARXIV.1707.06742

[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. https://doi.org/10.48550/ARXIV.1312.6034

[67] Hojun Son, Divya Krishnagiri, V. Swetha Jeganathan, and James Weiland. 2020. Crosswalk Guidance System for the Blind. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 3327–3330. https://doi.org/10.1109/EMBC44109.2020.9176623

[68] Joan Sosa-García and Francesca Odone. 2017. "Hands On" Visual Recognition for Visually Impaired Users. *ACM Trans. Access. Comput.* 10, 3, Article 8 (Aug. 2017), 30 pages. https://doi.org/10.1145/3060056

[69] Pierre Stock and Moustapha Cisse. 2018. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In *The European Conference on Computer Vision (ECCV)*.

[70] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 403–412.

[71] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. 2019. DEEP-HEAR: A Multimodal Subtitle Positioning System Dedicated to Deaf and Hearing-Impaired People. *IEEE Access* 7 (2019), 88150–88162. https://doi.org/10.1109/ACCESS.2019.2925806

[72] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. https://doi.org/10.1145/3441852.3471225

[73] David Touretzky. 2019. The AI4K12 Initiative: Developing National Guidelines for Teaching AI In K-12. https://github.com/touretzkyds/ai4k12/blob/master/documents/CSTA_2019_How_To_Teach_AI_Across_K-12.pdf.

[74] David Touretzky. 2020. The AI4K12 Initiative: Developing National Guidelines for Teaching AI In K-12. https://raw.githubusercontent.com/touretzkyds/ai4k12/master/documents/GlobalSWEdu2020_Touretzky.pdf.

[75] Henriikka Vartiainen, Matti Tedre, and Teemu Valtonen. 2020. Learning machine learning with very young children: Who is teaching whom? *International Journal of Child-Computer Interaction* 25 (Sept. 2020), 1–11. https://linkinghub.elsevier.com/retrieve/pii/S2212868920300155

[76] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3630–3638. http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf

[77] Vuzix. 2021. *Vuzix Blade Smart Glasses.* https://www.vuzix.com/products/blade-smart-glasses-upgraded

[78] Yutaro Yamanaka, Seita Kayukawa, Hironobu Takagi, Yuichi Nagaoka, Yoshimune Hiratsuka, and Satoshi Kurihara. 2022. One-Shot Wayfinding Method for Blind People via OCR and Arrow Analysis with a 360-Degree Smartphone Camera. In *Mobile and Ubiquitous Systems: Computing, Networking and Services*, Takahiro Hara and Hirozumi Yamaguchi (Eds.). Springer International Publishing, Cham, 150–168. https://doi.org/10.1007/978-3-030-94822-1_9

[79] Guangxiao Zhang, Zhuolin Jiang, and Larry S Davis. 2012. Online semi-supervised discriminative dictionary learning for sparse representation. In *Asian conference on computer vision*. Springer, 259–273. https://doi.org/10.1007/978-3-642-37331-2_20

[80] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. 2021. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731* (2021).

[81] Yuhang Zhao, Elizabeth Kupferstein, Brenda Veronica Castro, Steven Feiner, and Shiri Azenkot. 2019. Designing AR Visualizations to Facilitate Stair Navigation for People with Low Vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 387–402. https://doi.org/10.1145/3332165.3347906

[82] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173789

[83] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. *CoRR* abs/1801.05927 (2018). arXiv:1801.05927 http://arxiv.org/abs/1801.05927

[84] Zoom. 2022. *Video Conferencing, Cloud Phones, Webinars, Chat, Virtual Events.* https://zoom.us/