VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People

Danna Gurari[1], Qing Li[2], Chi Lin[1], Yinan Zhao[1], Anhong Guo[3], Abigale Stangl[4], Jeffrey P. Bigham[3]
[1] University of Texas at Austin [2] University of California, Los Angeles [3] Carnegie Mellon University [4] University of Colorado Boulder

Abstract

We introduce the first visual privacy dataset originating from people who are blind in order to better understand their privacy disclosures and to encourage the development of algorithms that can assist in preventing their unintended disclosures. It includes 8,862 regions showing private content across 5,537 images taken by blind people. Of these, 1,403 are paired with questions and 62% of those directly ask about the private content. Experiments demonstrate the utility of this data for predicting whether an image shows private information and whether a question asks about the private content in an image. The dataset is publicly-shared at `http://vizwiz.org/data/`.

1. Introduction

Mobile devices with built-in cameras have become ubiquitous. However, for people who are blind, using these devices to take and share pictures bares a grave risk of broadcasting private information [10, 12, 17, 44]. This is because blind people[1] by definition cannot see what is around them, and so cannot know what is in the field of view of their cameras. Still, many blind people share pictures they take in order to gain a transformative new ability to receive assistance in learning about their visual surroundings [7, 14, 15, 16, 18, 28, 31, 45, 55]. Some blind people also share their pictures on social media to socialize and express their creativity [8, 10, 26, 44]. And potentially many more of the 285 million people worldwide with visual impairments [38] would take and share pictures if given assurance that they could avoid privacy leaks [11]. Protecting private information is necessary to avoid the potential adverse social, professional, financial, and personal consequences to photographers (and bystanders) inflicted by privacy leaks.

While a natural step towards empowering blind people to protect private visual information is for the computer vision community to design algorithms to assist, a key obstacle is that none of the existing visual privacy datasets [21, 37, 36, 41, 52] needed to train algorithms are goal-oriented towards the images taken by and interests of people who are blind. Yet, our analysis shows that privacy leaks from this population are common; i.e., over 10% of more than 40,000 of their images contain private visual information. Moreover, our analysis suggests that over 50% of privacy leaks arise because people are explicitly compromising their privacy in exchange for assistance to learn about private visual information that is inaccessible to them (e.g., reading the number on a new credit card).

Our aim is to encourage the development of algorithms that can empower blind photographers to avoid inadvertently sharing private visual information. We begin by introducing the first visual privacy dataset originating from this population. The images were taken and shared by blind users of a visual question answering service [14]. For each image, we manually annotate private regions according to a taxonomy that represents privacy concerns relevant to their images, as summarized in Figure 1 (e.g., reading the results to a pregnancy test). We also annotate whether the private visual information is needed to answer the question asked by the user. These annotations serve as a critical foundation for designing algorithms that can decide (1) whether private information is in an image and (2) whether a question about an image asks about the private content in the image (a novel problem posed by our work). We benchmark numerous algorithms for both purposes. Our findings offer encouraging results that it is possible to automatically make both types of predictions while also demonstrating that the datasets are challenging for modern vision algorithms.

More generally, our work offers a new dataset challenge for training algorithms using large-scale, corrupted datasets. That is because the solution we propose for creating a large-scale, *publicly-available* visual privacy dataset is to remove private regions while preserving the context. Our empirical analysis of obfuscating the private content through various inpainting methods highlights (in)effective strategies for training algorithms using corrupted datasets.

[1] Currently, there is an international discussion on whether to use the phrase "blind person" versus "person who is blind" [1]. At present, both are accepted. For example, in the United States "person who is blind" is often preferred while in the United Kingdom "blind person" is preferred.

Figure 1. Examples of the types of private visual information found in images taken by blind people. Included are images that show common private objects (green box) as well as objects that commonly show private text (blue box). All private content is masked out (regions outlined in red) and replaced with automatically-generated fake content in order to safeguard the private information while preserving the context.

2. Related Work

Assistive Blind Photography. An increasing number of automated solutions are emerging that can assist blind people to take pictures. For example, many tools automatically guide users to improve the image focus, lighting, or composition in order to take a high quality picture [9, 15, 24, 26, 43, 54]. Other tools automatically notify users about what content is present (e.g., objects, text, generic descriptions) [6, 46], a valuable precursor for enabling users to decide if they are satisfied or should take another picture. While prior work demonstrates exciting progress, it does not yet explore the important problem of assisting users to protect private content. Yet, blind people have expressed concerns about inadvertently disclosing private information to the wrong people [10, 12, 17, 44] and the possible repercussions from privacy leaks are grave—e.g., identity theft, embarrassment, blackmail, legal liability. Accordingly, we introduce the first dataset challenge to encourage the development of algorithms that can alert this population when their pictures contain private information.

Visual Privacy Datasets. Large-scale datasets are often shared publicly to encourage the development of algorithms that automatically analyze visual information [19, 33, 39, 47]. Unfortunately, creating a large-scale collection of "private" images is inherently challenging since the very nature of such images means they are rarely shared. Still, several teams have successfully curated and shared private images that individuals posted for public viewing on photo-sharing websites (i.e., Flickr, Twitter) [36, 37, 52]. We, in contrast, curate images from people who agreed to the terms of the VizWiz mobile application that "Photos... may be... released as data sets... [after] personally-identifying information (photos, questions, etc) will be removed." As a result, we cannot release the private images as is. Other teams who faced a similar constraint addressed it by either (1) constraining data collection to consenting individuals (some who participated in staged private situations) [21, 41] or (2) releasing features describing private images (risking that future work may show how to recover the original content) [41]. We instead remove private regions and preserve their context in order to develop a large-scale visual privacy dataset that originates from a natural setting. This is the first visual privacy dataset originating from blind photographers and poses a new challenge of training algorithms to recognize private information using only its context. Experiments show the benefit of this dataset for training algorithms.

Visual Question Answering. Many large-scale visual question answering (VQA) datasets have been proposed to catalyze research on the VQA problem; e.g., [13, 22, 25, 27, 29, 34, 51]. However, no existing datasets include questions about private visual content. Yet, many people who are blind opt to ask questions about private visual content that is inaccessible to them; e.g., sharing a picture showing their prescription pills to learn what type of pills is in the bottle may be a lesser evil than accidentally taking the wrong pills. Accordingly, we pose a novel problem of predicting whether a question asks about private visual content, a critical precursor to deciding if the private content needs to be shared in order for a visual assistant to answer the question.

Taxonomy of Private Visual Information. A key challenge in designing a privacy dataset is establishing what is private in images. While legal and government entities offer laws and policies instructing how to protect privacy at large [20, 30, 35, 40, 42], their guidance leaves room for interpretation for images. Thus, researchers have proposed taxonomies based on people's stated privacy preferences both for images they take and images they see on social image sharing websites [32, 36, 37, 41, 52]. We propose the first taxonomy motivated by this population's images and report how often each privacy type arises in practice.

3. VizWiz-Priv Dataset

We now introduce "VizWiz-Priv" which consists of images taken by people who are blind. It builds off prior work that introduced a mobile phone application for enabling users to take pictures, ask questions about them, and have remote humans provide answers [14]. We created the dataset using 13,626 images taken by users who agreed to have their data shared anonymously. Note that this is a distinct set of images from the 31,173 that already are publicly available as part of the VizWiz-VQA dataset [25]. These image were excluded from VizWiz-VQA because they either lacked questions (and so are irrelevant for VQA) or were flagged by prior work [25] as containing private information and so unsuitable for public consumption. In what follows, we describe how we annotated private visual content, created the dataset, and then analyzed both.

3.1. Annotation of Private Visual Content

Privacy Taxonomy and Annotation Tool. We began by developing a taxonomy of visual privacy issues as well as an annotation tool for localizing private content. Initially, we designed a prototype for this purpose. Then, three trusted individuals who contributed to the design of the prototype independently used it to annotate private content in 120 randomly-selected images, and subsequently refined the taxonomy to resolve their annotation differences as well as the design of the tool to improve its usability.

Our final privacy taxonomy was designed to reflect the types of image content that introduce risks to people who are blind. Accordingly, the candidate categories came from existing taxonomies [21, 25, 37, 49] and was refined to reflect those detected in the 120 annotated images, including extra categories that reflect how people who are blind subject themselves to risk (e.g. sharing medical information on pill bottles) as well as potential stigmas. This resulted in a two-level taxonomy hierarchy. We chose as the top-level categories "objects" and "text" since those were most frequently observed. Object categories, in turn, consist of tattoos, pregnancy test results, and faces (represented directly as well as in framed pictures or reflected off shiny surfaces since these two related categories were also frequently observed). Private text categories consist of 14 objects on which private text is commonly located and are listed in Figure 1; e.g., prescription pills show people's names; letters show addresses; and credit cards provide access to people's money. Also included are two categories for both objects and text to reflect "suspicious" content for which it is hard to decipher if private information is present (e.g., in poor quality images or complex scenes) as well as an "other" category to capture private content not in our taxonomy.

The final annotation tool supported a user to locate image regions showing private content and assign a privacy category to each region. Specifically, a user traces the boundary of private information (text or object) by clicking on the image with a series of points that are connected with straight lines and clicking on the first point to complete the region. Once finished, the person selects the first-level and second-level privacy categories. A user repeats this two-step process for all private regions in an image before moving to the next image. For regions densely populated with more than five of the same private object (e.g., crowd of people), users were instructed to annotate it as one region. For images lacking private information, a user could select a button to mark the image safe for public consumption as is.

Annotation Collection. We implemented a multi-level review process. First, an in-house annotator annotated the privacy information in all images. Then, two domain experts who helped define the taxonomy and annotation tool reviewed each annotated image to correct any mistakes, including updating assigned privacy categories and improving/adding polygons. For the latter, a person could adjust a polygon's vertices by dragging and dropping them.

To assess the reliability of the annotation process, we measured the annotation agreement between the two domain experts who reviewed all annotations. They each independently annotated 250 randomly-selected images. We found that they agreed on whether private information is present for 90% of images, agreeing 115 images do not contain private information and 109 images contain private information. This finding suggests that the annotators largely shared a common understanding for what is private.

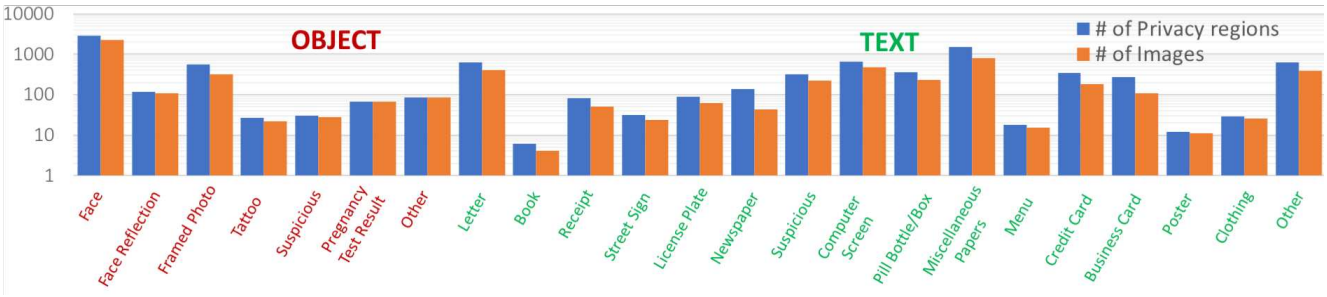

Figure 2. Frequency of disclosures for different privacy types with respect to number of images and number of private regions annotated. Results are shown on a logarithmic scale to illustrate the frequency of both common and rare causes of privacy disclosures.

	Private Objects					Private Text				
Type:	All	Face	Reflect	Photo	PregTest	All	MisPaper	ComScreen	Mail	Other
Area	159,881	166,169	80,185	99,289	53,134	99,259	82,066	77,507	68,516	88,977
$\frac{\textbf{Region Area}}{\textbf{Image Area}}$	0.11	0.11	0.05	0.07	0.03	0.06	0.05	0.05	0.05	0.04
Shape	0.82	0.85	0.81	0.75	0.67	0.55	0.54	0.56	0.57	0.54

Table 1. Mean value of properties describing private regions for different types of private information.

Analysis of Private Information. Using the annotated images, we then tallied how often each privacy type occurred across all images and private regions. Figure 2 shows the results, broken down with respect to the first-level categories ("object", "text") and 23 second-level categories.

Overall, 5,537 of the images were tagged as containing private information. When considering the larger collection of images taken with the VizWiz mobile phone application [14] (i.e., 13,626 + 31,173 [25] = 44,799), this means roughly 12% of all pictures taken by blind people show private content. This finding reveals that safeguarding private information is an important practical problem. We argue this problem is especially concerning since many visual assistance services still rely on humans [2, 3, 4, 7, 14, 23, 48].

A total of 8,862 private regions were identified across all private images. Slightly more of these regions were tagged as showing private text (i.e., 58%) than of private objects (i.e., 42%). Among the 5,151 text disclosures, they most commonly were found on miscellaneous papers (i.e., 30%) followed by computer screens (13%), letters (12%), and other objects (12%). Among the 3,711 object disclosures, the most common were faces (76%) followed by framed photographs (15%), face reflections (3%), and pregnancy test results (2%) respectively. These findings illustrate VizWiz-Priv offers a domain shift from the most similar privacy dataset [36]; [36], in contrast, covers only two of these eight most common categories: faces and letters/mail.

We also tallied across all 5,537 private images how many private regions and privacy types are in each image. Most commonly, one private region is detected per image (i.e., 67% of images), followed by two (i.e., 19% of images), three (i.e., 8%), and four or more (i.e., 6% of images). This finding directly influences our findings for the number of privacy categories per image with most commonly one type per image (i.e., 93%), followed by two (i.e., 6%), and at most three (i.e., <1%). This latter finding contrasts prior work which reported an average of 5.2 types per image for the VISPR dataset [37]. VizWiz-Priv averages 1.6 types per image. We hypothesize this discrepancy arises as a result of differences in the complexity of images, with VISPR more commonly biased towards complex scenes and VizWiz-Priv biased towards single object images.

We next characterized the appearance of private regions for different privacy categories. For each region, we computed its (1) area (i.e., absolute number of pixels in the region), (2) relative size (i.e., fraction of pixels in the image that belong to the region), and (3) circularity (i.e., ratio of its area A to a circle with the same perimeter P ($\frac{4\pi A}{P^2}$)). Table 1 shows the resulting mean values with respect to the top-level categories ("object", "text") and their most common second-level categories. We observe that private objects tend to be more circular than private text (i.e., 0.82 versus 0.55), possibly capturing the oval shape common for faces versus rectangular shape common for text. Also shown is that a region's relative size is almost two times larger for private objects (i.e., 11% of image) than private text (i.e., 6% of image). Moreover, text categories tend to be consistent in their relative size (i.e., 4% to 5%) compared to object categories (i.e., 3% to 11%); e.g., pregnancy tests occupy on average 3% of the image (with ∼53,134 pixels) whereas faces occupy on average 11% (with ∼166,169 pixels). These findings reveal different types of private content exhibit different visual biases, a valuable precursor for algorithms learning to recognize and distinguish between them.

Dataset	Image Source	# Images	Annotation	Taxonomy	Public Content
Campus Face Set [21]	Moving Vehicle	2,176	Rectangle	Faces	Images (of Staged Actors)
PicAlert [52]	Flickr	4,701	Image Label	Public/Private	Images
YourAlert [41]	Social Network	1,511	Image Label	Public/Private	Image Features
VISPR [37]	Flickr, Twitter	∼12,000	Image Label	68 Categories	Images
Redactions [36]	Flickr, Twitter	8,473	Polygons + Labels	24 Categories	Images + Masks + Labels
Ours: VizWiz-Priv	Blind People	5,537	Polygons + Labels	23 Categories	Masked Images + Labels

Table 2. Comparison of five image privacy datasets and our VizWiz-Priv dataset. (Note: "# Images" indicates the number of private images)

3.2. Dataset Creation

Since the privacy concerns we are addressing prohibit us from releasing publicly the private content in images, we opted to release only the context around it. Thus, a primary goal in designing the dataset was to minimize the possibility that algorithms will learn to detect artifacts around removed regions as predictive about private content. In what follows, we describe how we created VizWiz-Priv with this aim in mind as well as our analysis of the dataset.

Removing Regions from Private & Non-Private Images. Our aim is to ensure private and non-private images share similar artifacts. Thus, in addition to masking out the private regions from all 5,537 private images, we randomly applied those same private regions to the remaining 8,093 non-private images to determine what content to remove. Doing so ensures the same shape and location statistics of the removed regions are applied to both the private and non-private images. While this approach does have a shortcoming that the masked out regions in the non-private images may not cover meaningful objects or chunks of text, it will be shown in Section 5 that an algorithm can still learn cues that are predictive of private information.

Inpainting to Replace Removed Regions. Next, we synthesized content to fill in regions removed from all images in attempt to make them look more realistic for algorithm training as well as for human review in the publicly-released dataset. We employed the publicly-available code for the state-of-the-art image inpainting system [50]. This approach explicitly utilizes surrounding image context to decide how to synthesize novel image structures in a removed region. We also created another version where we replace all hole pixels with the mean value from ImageNet.

Dataset Comparison. Table 2 illustrates how VizWiz-Priv compares to existing privacy datasets.

One key distinction of VizWiz-Priv is the method for publicly-releasing the data, which stems from the use case which led to the collection of the images. While most teams could release images as is, since images came from individuals who already posted them for public viewing [36, 37, 52], two teams shared our constraint that no private content could be released. These teams addressed this constraint by either employing "actors" who consented to participate in a similar staged private situation [21] or releasing features describing private images (risking that future work may show how to recover the original content) [41]. We instead removed private regions while preserving their context. Thus, VizWiz-Priv poses a new challenge of how to successfully train privacy detection algorithms using only the context in which private information is located.

Another key distinction is that VizWiz-Priv is the first privacy dataset originating from blind photographers trying to learn about their visual surroundings. Consequently, unlike images in existing datasets, many images are poor quality (e.g., blurry, out of focus) since the photographers cannot verify their quality. Additionally, much of the content is concentrated around private information that is currently inaccessible since blind people commonly shared this information in exchange for assistance to learn about it (see Section 4); e.g., pregnancy test results, prescription pills, business cards, and street signs. A final bias of VizWiz-Priv is that many images show indoor, home scenes and the unintentional privacy leaks that can arise in this setting, such as reflections of people's faces on computer/television screens, personal documents sitting on counter-tops, and personal photographs strewn along the walls.

4. Visual Question Answering (VQA)

An important consideration when empowering blind people to protect their privacy is to avoid impeding their ability to solicit the information they seek. Specifically, in the VQA setting, a naive solution to enhance privacy over the status quo of sharing every visual question with remote visual assistants (e.g., for services such as VizWiz[14] and BeSpecular [3]) is to instruct the photographer to retake the picture (or automatically mask out the private information) whenever private information is detected in the image. Unfortunately, this would be inappropriate when a person is trying to learn about the private information (e.g., to learn what type of pills are in a bottle). Instead, a user would benefit from taking another picture before sharing the visual question *only* for unintentional privacy leaks. Accordingly, we now describe our preparation of a dataset for training

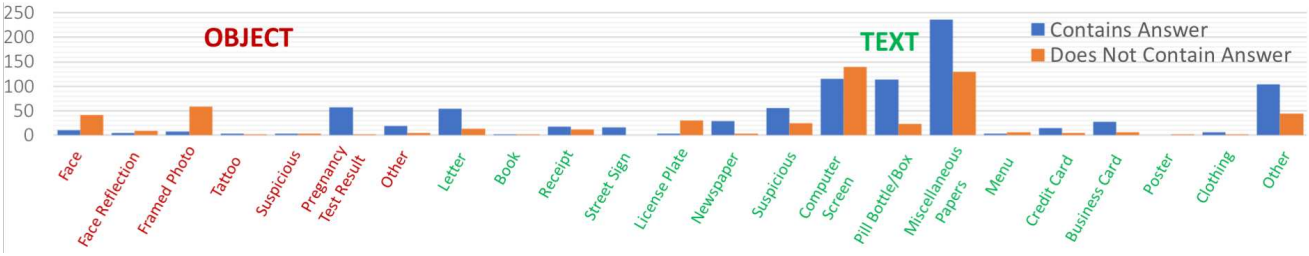

Figure 3. Frequency of disclosures per privacy type with respect to the number of images for which a person asked about private content.

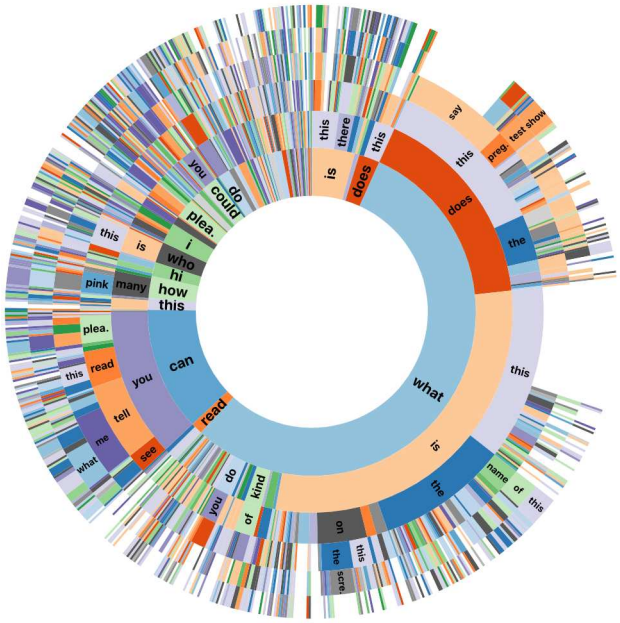

Figure 4. Frequency of questions beginning with different words/phrases for the 875 instances in VizWiz-Priv where the answers reside in private regions of the images. The innermost ring represents the first word and each subsequent ring represents a subsequent word up to six words in the question. The arc size is proportional to the number of questions containing that word/phrase.

algorithms to automatically determine whether the question asks about the private content. We use the 2,685 images in VizWiz-Priv that also contain questions. Of these, 1,403 contain private images.

Visual Question Pre-Processing. For each visual question, we followed prior work's [25] pre-processing steps. Specifically, we re-saved images to remove personally-identifying metadata, transcribed audio-recorded questions to remove people's voices, and spell-checked questions[2].

VizWiz-Priv-VQA. Next, we quantified how often private visual information contains the answer to a visual question. To do so, the same three in-house annotators who developed VizWiz-Priv reviewed the 1,403 visual questions that contain private images and indicated "Yes" if the question would become unanswerable by removing the private visual content or "No" otherwise. Each visual question was assigned the majority vote label from the annotators.

We found 62% (i.e., 875) of the 1,403 visual questions asked about private visual content. Within the larger context of all 33,858 visual questions asked by VizWiz users (i.e., 2,685 + 31,173 [25] visual questions), this means that more than 1 of every 40 visual questions arose because blind people were compromising their privacy in exchange for visual assistance. Moreover, this statistic is likely a lower bound of the actual visual privacy assistive needs of this population since many people avoid sharing private information rather than accepting the risks of sharing [11]. Our finding underscores the importance of designing algorithms that can answer questions about private visual information.

We next quantified the tendency for each type of private information to be in the images based on whether people were explicitly asking about private content[3]. Figure 3 shows the results. We found that a person was asking about private content for most images showing pregnancy test results (i.e., 58/59=98%), pill bottles/boxes (i.e., 114/137=83%), letters (i.e., 55/68=81%), street signs (i.e., 16/16=100%), credit cards (i.e., 15/20=75%), and more. Numerous privacy categories also often occurred when private content was unnecessarily captured in images. For example, the private content could be safely removed without impacting the ability to answer the visual question for roughly 81% of faces, 89% of framed pictures, and 88% of license plates. As shown, a person's intentions to share private information can be correlated to the type of private information present in an image.

We visualize the questions people asked for the 875 visual questions which asked about the private content using a sunburst diagram, shown in Figure 4. This illustrates the frequency that the questions begin with different words/phrases. When comparing the questions reported in prior work about non-private questions [25] to these private questions, we observe great similarity. For example, both collections share a rich diversity of first words and are sim-

[2]We collected 10 answers per visual question for the 2,685 visual questions in order to contribute an 8% increase to the size of VizWiz-VQA [25].

[3]Since some images show multiple private regions, this analysis also reveals correlation of extra privacy type(s) present in an image when a person asks a question about a different type of private content that is present.

ilar in question length. However, we observe a shift in the frequency of particular questions; e.g., the recognition question of "What is this?" occurs less often (i.e., 2.5 times less often with 5.8% versus 14.6% in [25]), whereas the reading question of "What does this say?" arises more regularly (i.e., 4 times more often with 4% of questions versus 1% in [25]). We hypothesize this shift away from object recognition and towards reading exemplifies the broader tendency shown in Figure 3 that people more often intentionally capture pictures with private text (i.e., 65% of text disclosures) than with private objects (i.e., 47% of object disclosures).

5. Algorithm Benchmarking

We now describe two studies conducted for the tasks of predicting if (1) private content is present in a given image and (2) a visual question asks about private content.

5.1. Private Visual Information Recognition

Dataset. We divided all 13,626 VizWiz-Priv images into approximately a 65-10-25 split resulting in 8,825, 1,370, and 3,431 images in the training, validation, and test sets. We perform a stratified split on the private images with respect to privacy categories in an attempt to include a proportional number of each privacy type in each split.

Methods. As done for the state-of-art visual privacy recognition algorithm [37], we also fine-tuned ResNet-50. We benchmarked 10 variants. Four were fine-tuned to the training datasets from VISPR [37], VizWiz-Priv with the hole inpaintings (`VizWiz-Priv`), VizWiz-Priv with the ImageNet mean assigned to hole pixels (`VizWiz-Priv-HoleMean`), and the original VizWiz-Priv images (`VizWiz-Priv-Uncorrupted`) respectively. Another three were the VISPR-trained method [37] fine-tuned to each of the following three datasets: `VizWiz-Priv`, `VizWiz-Priv-HoleMean`, and `VizWiz-Priv-Uncorrupted`. Finally, three were fine-tuned to the combination of the VISPR dataset with each of the above three datasets. Each method was fine-tuned using the Adam solver with a batch size of 128 and fixed learning rate of 0.001, employing dropout and batch normalization during training, and training for five epochs.

Evaluation Metrics. We evaluated each method using a precision-recall (PR) curve and the average precision (AP).

Results on Uncorrupted VizWiz-Priv Images. In Figure 5, we report the performance of each method when evaluated on the original, uncorrupted VizWiz-Priv test images in order to demonstrate their utility in a practical setting.

We found that the state-of-art model [37] generalizes well to the uncorrupted VizWiz images; i.e., AP score of 77.97%. This offers an exciting finding since the model was not trained on images coming from blind photographers.

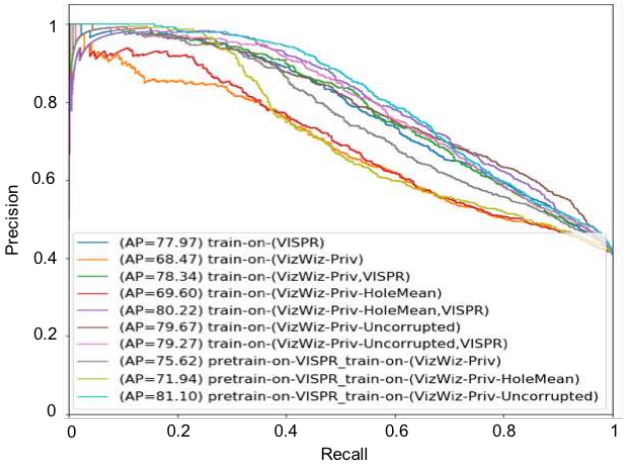

Figure 5. Precision-recall curves and AP scores for privacy detection algorithms evaluated on the uncorrupted VizWiz-Priv test set.

We observe considerable improvements when training with the original VizWiz-Priv images. In comparison to the above VISPR-trained model, we observe a 2% boost when training the same architecture instead with the uncorrupted VizWiz-Priv images and a 3% boost when fine-tuning the VISPR-trained model with those images. These findings reinforce the well-known benefit that training on data matching the test distribution yields improved performance.

Given the privacy concerns motivating this work, it also is valuable to analyze the methods trained only with the versions of VizWiz-Priv that will be publicly-available; i.e., those where private content is masked out of the images (`VizWiz-Priv`, `VizWiz-Priv-HoleMean`). When training directly with this data, we observe almost a 13% performance drop from today's state-of-art model; i.e., 77.97% versus 68.47% and 69.6%. We also observe inferior performance when fine-tuning the VISPR-trained model to these datasets; i.e., AP score drops from 77.97% to 75.62% and 71.94%. We attribute these declines to the inadequacies of the hole-filling methods; as exemplified in Figure 1, the hole-filling algorithm can introduce visual artifacts that algorithms may be learning to model. Yet, interestingly, we observe a boost in predictive performance when training using the VISPR images jointly with both public versions of VizWiz-Priv; i.e., AP score improves from 77.97% to 80.22% and 78.34%. In fact, training with both uncorrupted images (VISPR) and corrupted images (VizWiz-Priv) outperforms training with either dataset alone. We hypothesize that training with both datasets helps the model to learn to ignore the hole-filling artifacts isolated to VizWiz-Priv while providing richer training data needed to successfully learn cues that are predictive across both datasets.

Results on VISPR Images. We also examined the performance of each baseline on the test set for the existing state-of-art visual privacy dataset—VISPR [37], which is

a collection of images scraped from online photo-sharing websites. The resulting AP scores for the ten methods range from 88.08% (for fine-tuning to VizWiz-Priv-uncorrupted) to 96.78% (for [37]). This highlights that VizWiz-Priv offers a more difficult problem than VISPR, with the top-performing algorithms achieving an AP score of 96.8% on VISPR versus 81.1% on VizWiz-Priv-Uncorrupted.

VizWiz-Priv versus VizWiz-Priv-Uncorrupted Results. We also evaluated how well the uncorrupted version of VizWiz-Priv represents the publicly-available versions. We found a high correlation between predicted scores on the uncorrupted VizWiz-Priv images and hole-filled VizWiz-Priv images; i.e., when computing the Pearson correlation coefficient, scores ranged from 0.70 to 0.89. While imperfect, this test set offers a reasonable privacy-free substitute to benchmark algorithm performance in a reproducible way.

5.2. (Un)intentional Privacy Leak Recognition

We now evaluate methods for a novel binary classification problem of predicting whether a given visual question asks about private visual information in order to distinguish between intentional and unintentional privacy disclosures.

Dataset. We performed a stratified split on the 2,685 visual questions in order to preserve the proportions of private versus non-private images. This resulted in 2,148 and 537 visual questions in the training and test sets respectively.

Methods. We benchmarked ten methods. We evaluated a `Status Quo` predictor which returns a random value to reflect the best a system can achieve today; i.e., guessing. Also included are two related privacy detection algorithms from the previous section; i.e., `Priv-Detection [37]` and `Priv-Detection-V&V` (training on both VizWiz-Priv and VISPR). We also predicted directly from questions (i.e., `Q`), encoding each as a 300-dimensional GloVe word-embedding and training a single-layer gated recurrent unit (GRU) network with 300 hidden states. We additionally predicted directly from images, encoding each image with ResNet-50 and training three variants: using the images as is (i.e., `I-original`), with private regions replaced by mean values (i.e., `I-hole-mean`), and with private regions replaced by the hole-filling algorithm (i.e., `I-hole-inpaint`). Finally, we investigated three models that combine the question with each image variant.

Evaluation Metrics. We assessed the performance of each method using a precision-recall curve and its AP score.

Results. Results are shown in Figure 6. As observed, it is possible to predict whether a visual question asks about private content directly from the visual question, despite the significant variety of privacy types and question types. For example, all `Q+I` methods outperform the status quo approach by at least 30 percentage points. Our findings also demonstrate the value in directly learning for the task rather than relying on predictors for the related task of detecting private visual information; i.e., AP improves from below 40% to above 60% for all `Q + I` methods.

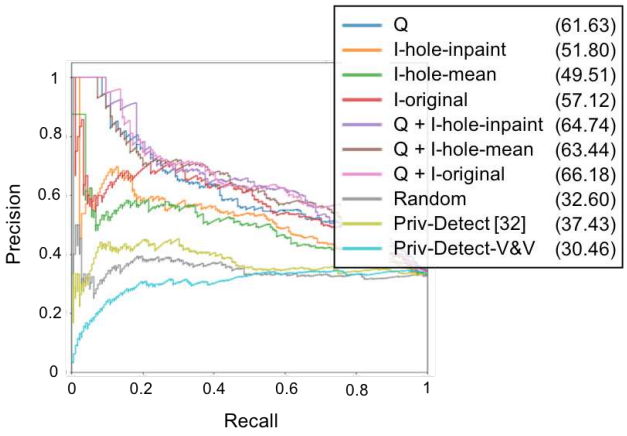

Figure 6. PR curves and AP scores for algorithms that predict whether a visual question asks about private visual content.

We observe that the question and image each offer valuable and complementary clues regarding whether a visual question asks about private information. Specifically, while both the question (`Q`) and image (`I-original`) are predictive alone (i.e., 30.25 and 24.9 percentage point boost over the status quo respectively), we observe a further 3.5 percentage point boost when using these features jointly.

Our findings also highlight the utility of images with synthesized content for training algorithms. We observe that training using the context of private content without the private content itself improves accuracy by approximately 17 percentage points compared to the status quo approach. This benefit arises whether filling holes using the image mean or state-of-art hole-filling algorithm. These findings highlight it is possible to develop privacy-based algorithms without having access to the private content.

6. Conclusions

We proposed the first dataset and study that reveals visual privacy issues faced by people who are blind who are trying to learn about their physical surroundings. Experiments demonstrated its benefit for teaching algorithms to predict the presence and purpose of private information. Valuable future work includes (1) detecting private content to support redacting it [36] (a more challenging problem than recognition), (2) improving hole-filling algorithms to replace private content (e.g., [53]), (3) expanding the privacy taxonomy (e.g., with scene types and actions), and (4) adding images from wearable cameras (e.g., [4, 5]).

Acknowledgements. We thank the anonymous reviewers for their valuable feedback. This work is supported in part by National Science Foundation funding (IIS-1755593) as well as gifts from Adobe and Microsoft to Danna Gurari.

References

[1] Accessible Writing Guide — SIGACCESS. 1

[2] Be My Eyes - Bringing sight to blind and low-vision people. https://www.bemyeyes.com/. 4

[3] BeSpecular. https://www.bespecular.com. 4, 5

[4] Home - Aira : Aira. https://aira.io/. 4, 8

[5] Orcam. https://www.orcam.com/en/. 8

[6] Seeing AI — Talking camera app for those with a visual impairment. https://www.microsoft.com/en-us/seeing-ai. 2

[7] TapTapSee - Blind and Visually Impaired Assistive Technology - powered by the CloudSight.ai Image Recognition API. https://taptapseeapp.com/. 1, 4

[8] Tommy Edison (@blindfilmcritic) • Instagram photos and videos. https://www.instagram.com/blindfilmcritic/. 1

[9] D. Adams, L. Morales, and S. Kurniawan. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, page 25. ACM, 2013. 2

[10] T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532. ACM, 2015. 1, 2

[11] T. Ahmed, R. Hoyle, P. Shaffer, K. Connelly, D. Crandall, and A. Kapadia. Understanding the Physical Safety, Security, and Privacy Concerns of People with Visual Impairments. *IEEE Internet Computing*, 21(3):56–63, 2017. 1, 6

[12] T. Ahmed, P. Shaffer, K. Connelly, D. Crandall, and A. Kapadia. Addressing physical safety, security, and privacy for people with visual impairments. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS'16)*, pages 341–354, 2016. 1, 2

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 3

[14] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, and S. White. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, pages 333–342. ACM, 2010. 1, 3, 4, 5

[15] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference On*, pages 65–72. IEEE, 2010. 1, 2

[16] E. L. Brady, Y. Zhong, M. R. Morris, and J. P. Bigham. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1225–1236. ACM, 2013. 1

[17] S. M. Branham, A. Abdolrahmani, W. Easley, M. Scheuerman, E. Ronquillo, and A. Hurst. Is Someone There? Do They Have a Gun: How Visual Information about Others Can Improve Personal Safety Management for Blind Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 260–269. ACM, 2017. 1, 2

[18] M. A. Burton, E. Brady, R. Brewer, C. Neylan, J. P. Bigham, and A. Hurst. Crowdsourcing subjective fashion advice using VizWiz: Challenges and opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 135–142. ACM, 2012. 1

[19] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[20] U. S. Congress. The Privacy Act of 1974. *Public Law*, 88, 1974. 3

[21] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in Google Street View. In *ICCV*, pages 2373–2380, 2009. 1, 2, 3, 5

[22] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, volume 1, page 3, 2017. 3

[23] D. Gurari and K. Grauman. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3511–3522. ACM, 2017. 4

[24] D. Gurari, K. He, B. Xiong, J. Zhang, M. Sameki, S. D. Jain, S. Sclaroff, M. Betke, and K. Grauman. Predicting Foreground Object Ambiguity and Efficiently Crowdsourcing the Segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018. 2

[25] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 3, 4, 6, 7

[26] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 203–210. ACM, 2011. 1, 2

[27] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference On*, pages 1988–1997. IEEE, 2017. 3

[28] H. Kacorri, K. M. Kitani, J. P. Bigham, and C. Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5839–5849. ACM, 2017. 1

[29] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference On*, pages 1983–1991. IEEE, 2017. 3

[30] L. D. Koontz. *Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information*. DIANE Publishing, 2008. 3

[31] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 18. ACM, 2013. 1

[32] X. Li, D. Li, Z. Yang, and W. Chen. A Patch-Based Saliency Detection Method for Assessing the Visual Privacy Levels of Objects in Photos. *IEEE Access*, 5:24332–24343, 2017. 3

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2

[34] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014. 3

[35] E. McCallister. *Guide to Protecting the Confidentiality of Personally Identifiable Information*. Diane Publishing, 2010. 3

[36] T. Orekondy, M. Fritz, and B. Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 8

[37] T. Orekondy, B. Schiele, and M. Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Computer Vision (ICCV), 2017 IEEE International Conference On*, pages 3706–3715. IEEE, 2017. 1, 2, 3, 4, 5, 7, 8

[38] W. H. Organization. *Global Data on Visual Impairments 2010. Geneva: World Health Organization, 2012*. 1

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[40] D. J. Solove. A taxonomy of privacy. *U. Pa. L. Rev.*, 154:477, 2005. 3

[41] E. Spyromitros-Xioufis, S. Papadopoulos, A. Popescu, and Y. Kompatsiaris. Personalized privacy-aware image classification. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 71–78. ACM, 2016. 1, 2, 3, 5

[42] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling. PrivacEye: Privacy-Preserving First-Person Vision Using Image Features and Eye Movement Analysis. *arXiv preprint arXiv:1801.04457*, 2018. 3

[43] M. Vázquez and A. Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(5):25, 2014. 2

[44] V. Voykinska, S. Azenkot, S. Wu, and G. Leshed. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1584–1595. ACM, 2016. 1, 2

[45] M. Weiss, M. Luck, R. Girgis, C. Pal, and J. Cohen. A Survey of Mobile Computing for the Visually Impaired. *arXiv preprint arXiv:1811.10120*, 2018. 1

[46] S. Wu, J. Wieland, O. Farivar, and J. Schiller. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *CSCW*, pages 1180–1192, 2017. 2

[47] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference On*, pages 3485–3492. IEEE, 2010. 2

[48] C.-J. Yang, K. Grauman, and D. Gurari. Visual Question Answer Diversity. In *HCOMP*, pages 184–192, 2018. 4

[49] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan. Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-Grained Privacy Settings for Social Image Sharing. *IEEE Transactions on Information Forensics and Security*, 13(5):1317–1332, 2018. 3

[50] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv preprint*, 2018. 5

[51] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the Ieee International Conference on Computer Vision*, pages 2461–2469, 2015. 3

[52] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–44. ACM, 2012. 1, 2, 3, 5

[53] Y. Zhao, B. Price, S. Cohen, and D. Gurari. Guided image inpainting: Replacing an image region by pulling content from another image. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1514–1523. IEEE, 2019. 8

[54] Y. Zhong, P. J. Garrigues, and J. P. Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 20. ACM, 2013. 2

[55] Y. Zhong, W. S. Lasecki, E. Brady, and J. P. Bigham. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2353–2362. ACM, 2015. 1