Assessing Image Quality Issues for Real-World Problems

Tai-Yin Chiu, Yinan Zhao, Danna Gurari

University of Texas at Austin

Abstract

We introduce a new large-scale dataset that links the assessment of image quality issues to two practical vision tasks: image captioning and visual question answering. First, we identify for 39,181 images taken by people who are blind whether each is sufficient quality to recognize the content as well as what quality flaws are observed from six options. These labels serve as a critical foundation for us to make the following contributions: (1) a new problem and algorithms for deciding whether an image is insufficient quality to recognize the content and so not captionable, (2) a new problem and algorithms for deciding which of six quality flaws an image contains, (3) a new problem and algorithms for deciding whether a visual question is unanswerable due to unrecognizable content versus the content of interest being missing from the field of view, and (4) a novel application of more efficiently creating a large-scale image captioning dataset by automatically deciding whether an image is insufficient quality and so should not be captioned. We publicly-share our datasets and code to facilitate future extensions of this work: `https://vizwiz.org`.

1. Introduction

Low-quality images are an inevitable, intermittent reality for many real-world, computer vision applications. At one extreme, they can be life threatening, such as when they impede the ability of autonomous vehicles [60] and traffic controllers [30] to safely navigate environments. In other cases, they can serve as irritants when they convey a negative impression to the viewing audiences, such as on social media or dating websites.

Despite that low-quality images often emerge in practical settings, there has largely been a disconnect between research aimed at recognizing quality issues and research aimed at performing downstream vision tasks. For researchers focused on uncovering what quality issues are observed in an image, their progress largely has grown from artificially-constructed settings where they train and evaluate algorithms on publicly-available datasets that were constructed by distorting high quality images to simulate quality issues (e.g., using JPEG compression or Gaussian blur) [41, 48, 12, 21, 37, 36, 25, 31]. Yet, these contrived environments typically lack sufficient sophistication to capture the plethora of factors that contribute to quality issues in natural settings (e.g., camera hardware, lighting, camera shake, scene obstructions). Moreover, the quality issues are detangled from whether they relate to the ability to complete specific vision tasks. As for researchers focusing on specific tasks, much of their progress has developed from environments that lack low-quality images. That is because the creators of popular publicly-available datasets that support the development of such algorithms typically included a step to filter out any candidate images that are deemed insufficient quality for the final dataset [11, 14, 23, 9, 53, 28, 59]. Consequently, such datasets lack data that would enable training algorithms to identify when images are of insufficient quality to complete a given task.

Motivated by the aim to tie the assessment of image quality to practical vision tasks, we introduce a new image quality assessment (IQA) dataset that emerges from a real use case. Specifically, our dataset is built around 39,181 images that were taken by people who are blind who were authentically trying to learn about images they took using the VizWiz mobile phone application [5]. Of these images, 17% were submitted to collect image captions from remote humans. The remaining 83% were submitted with a question to collect answers to their visual questions. As discussed in prior work [7, 17], users submitted these images and visual questions (i.e., images with questions) to overcome real visual challenges that they faced in their daily lives. They typically waited nearly two minutes to receive a response from the remote humans [5]. For each image, we asked crowdworkers to either supply a caption describing it or clarify that the quality issues are too severe for them to be able to create a caption. We call this task the *unrecognizability classification task*. We also ask crowdworkers to label each image with quality flaws that are more traditionally discussed in the literature [7, 12]: blur, overexposure (bright), underexposure (dark), improper framing, obstructions, and rotated views. We call this task the *quality flaws classification task*. Examples of resulting labeled images in our dataset are shown in Figure 1. Altogether, we call this

| No Flaws (NON) | Blur (BLR) | Bright (BRT) | Dark (DRK) | Framing (FRM) | Obscured (OBS) | Rotation (ROT) | Unrecognizable |

A dry golden leaf on a brownish red background. | A gold thermostat or compass attached to a piece of wood. | A black Acer brand laptop with monitor off. | A person holding a Chap Stick brand lip balm. | A hand holding a blue Kraft Foods jar. | A coin placed on a left hand or palm. | A green bag of potato chips from the store. | Quality issues are too severe to recognize visual content.

Quality Flaws　　　　　　　　　　　　　　　　　　Not Captionable

Figure 1: We introduce a new image quality assessment dataset which we call VizWiz-QualityIssues. Shown are examples of the taxonomy of labels, which ranges from no quality issues to six quality flaws to unrecognizable/uncaptionable images. Images can manifest different combinations of the above labels, for instance the unrecognizable image is also labeled as suffering from image blur and poor framing.

dataset VizWiz-QualityIssues.

We then demonstrate the value of this new dataset for several new purposes. First, we introduce a novel problem and algorithms for predicting whether an image is sufficient quality to be captioned (Section 4). This can be of immediate use to blind photographers, who otherwise must wait nearly two minutes to learn their image is unsuitable quality for image captioning. We next conduct experiments to demonstrate an additional benefit of this prediction system for creating large-scale image captioning datasets with less wasted human effort (Section 4.3). Finally, we introduce a novel problem and algorithms that inform a user who submits a novel visual question whether it can be answered, cannot be answered because the image content is unrecognizable, or cannot be answered because the image content is missing from the image (Section 5). This too can be of immediate benefit to blind photographers by enabling them to both fail fast and gain valuable insight into how to update the visual question to make it become answerable.

More generally, our work underscores the importance of defining quality within the context of specific tasks. We expect our work can generalize to related vision tasks such as object recognition, scene classification, and video analysis.

2. Related Work

Image Quality Datasets. A number of image quality datasets exist to support the development of image quality assessment (IQA) algorithms, including LIVE [41, 48], LIVE MD [21], TID2008 [37], TID2013 [36], CSIQ [25], Waterloo Exploration [31], and ESPL-LIVE[24]. A commonality across most such datasets is that they originate from high quality images that were artificially distorted to introduce image quality issues. For example, LIVE [12] consists of 779 distorted images, which are derived by applying five different types of distortions at numerous distor-

tion levels to 29 high-quality images. Yet, image quality issues that arise in real-world settings exhibit distinct appearances than those that are found by simulating distortions to high-quality images. Accordingly, our work complements recent efforts to create large-scale datasets that flag quality issues in natural images [12]. However, our dataset is considerably larger, offering approximately a 19-fold increase in the number of naturally distorted images; i.e., 20,244 in our dataset versus 1,162 images for [12]. In addition, while [12] assigns a single quality score to each image to capture any of a wide array of image quality issues, our work instead focuses on recognizing the presence of each distinct quality issue and assessing the impact of the quality issues on the real application needs of real users.

Image Quality Assessment. Our work also relates to the literature that introduces methods for assessing the quality of images. One body of work assumes that developers have access to a high-quality version of each novel image, whether partially or completely. For example, distorted images are evaluated against original, intact images for full-reference IQA algorithms [48, 50, 57, 41, 25, 6, 39] and distorted images are evaluated against partial information about the original, intact images for reduced-reference IQA algorithms [49, 26, 47, 42, 38, 32, 51]. Since our natural setting inherently limits us from having access to original, intact images, our work instead aligns with the second body of work which is built around the assumption that no original, reference image is available; i.e., no-reference IQA (NR-IQA). NR-IQA algorithms instead predict a quality score for each novel image [33, 22, 47, 56, 55, 29, 43, 6, 44]. While many algorithms have been introduced for this purpose, our analysis of five popular NR-IQA models (i.e., BRISQUE [33], NIQE [34], CNN-NRIQA [22], DNN-NRIQA [6], and NIMA [44]) demonstrates that they are inadequate for our novel task of assessing which images

are unrecognizable and so cannot be captioned (discussed in Section 4). Accordingly, we introduce new algorithms for this purpose, and demonstrate their advantage.

Efficient Creation of Large-Scale Vision Datasets. Progress in the vision community has largely been measured and accelerated by the creation of large-scale vision datasets over the past 20 years. Typically, researchers have scraped images for such datasets from online image search databases [11, 14, 23, 9, 53, 28, 59]. In doing so, they typically curate a large collection of high-quality images, since such images first passed uploaders' assessment that they are of sufficient quality to be shared publicly. In contrast, when employing images captured "in the wild," it can be a costly, time-consuming process to identify and remove images with unrecognizable content. Accordingly, we quantify the cost of this problem, introduce a novel problem and algorithms for deciphering when image content would be unrecognizable to a human and so should be discarded, and demonstrate the benefit of such solutions for more efficiently creating a large-scale image captioning dataset.

Assistive Technology for Blind Photographers. Our work relates to the literature about technology for assisting people who are blind to take high-quality pictures [1, 5, 20, 45, 58]. Already, existing solutions can assist photographers in improving the image focus [1], lighting [5], and composition [20, 45, 58]. Additionally, algorithms can inform photographers whether their questions about their images can be answered [17] and why crowds struggle to provide answers [4, 15]. Complementing prior work, we introduce a suite of new AI problems and solutions for offering more fine-grained guidance when alerting blind photographers about what image quality issue(s) are observed. Specifically, we introduce novel problems of (1) recognizing whether image content can be recognized (and so captioned) and (2) deciphering when a question about an image can be answered, cannot be answered because the image content is unrecognizable, or cannot be answered because the content of interest is missing from the image.

3. VizWiz-QualityIssues

We now describe our creation of a large-scale, human-labeled dataset to support the development of algorithms that can assess the quality of images. We focus on a real use case that is prone to image quality issues. Specifically, we build off of 39,181 publicly-available images [16, 17] that originate from blind photographers who each submitted an image with, optionally, a question to the VizWiz mobile phone application [5] in order to receive descriptions of the image from remote humans. Since blind photographers are unable to verify the quality of the images they take, the dataset exemplifies the large diversity of quality issues that occur naturally in practice. We describe below how we create and analyze our new dataset.

3.1. Creation of the Dataset

We scoped our dataset around quality issues that impede people who are blind in their daily lives. Specifically, a clear, resounding message is that people who are blind need assistance in taking images that are sufficiently high-quality that sighted people are able to either describe them or answer questions about them [5, 7].

Quality Issues Taxonomy. One quality issue label we assess is whether *image content is sufficiently recognizable* for sighted people to caption the images. We also label numerous *quality flaws* to situate our work in relation to other papers that similarly focus on assessing image quality issues [7, 12]. Specifically, we include the following categories: blur (is the image blurry?), bright (is the image too bright?), dark (is the image too dark?), obstruction (is the scene obscured by the photographer's finger over the lens, or another unintended object?), framing (are parts of necessary items missing from the image?), rotation (does the image need to be rotated for proper viewing?), other, and no issues (there are no quality issues in the image).

Image Labeling Task. To efficiently label all images, we designed our task to run on the crowdsourcing platform Amazon Mechanical Turk. The task interface showed an image on the left half and the instructions with user-entry fields on the right half. First, the crowdworker was instructed to either describe the image in one sentence or click a button to flag the image as being insufficient quality to recognize the content (and so not captionable). When the button was clicked, the image description was automatically populated with the following text: "Quality issues are too severe to recognize the visual content." Next, the crowdworker was instructed to select all image quality flaws from a pre-defined list that are observed. Shown were the six reasons identified above, as well as Other (OTH) linked to a free-entry text-box so other flaws could be described and None (NON) so crowd workers could specify the image had no quality flaws. The interface enabled workers to adjust their view of the image, using the toolbar to zoom in, zoom out, pan around, or rotate the image if needed. To encourage higher quality results, the interface prevented a user from completing the task until a complete sentence was provided and at least one option from the "image quality flaw" options was chosen. A screen shot of the user interface is shown in the Supplementary Materials.

Crowdsourcing Labels. To support the collection of high quality labels, we only accepted crowdworkers who previously had completed over 500 HITs with at least a 95% acceptance rate. Also, we collected redundant results. Specif-

ically, we recruited five crowdworkers to label each image. We deemed a label as valid only if at least two crowdworkers chose that label.

3.2. Characterization of the Dataset

Prevalence of Quality Issues. We first examine the frequency at which images taken by people who are blind suffer from the various quality issues to identify the (un)common reasons. To do so, we tally how often unrecognizable images and each quality-flaw arise.

Roughly half of the images suffer from image quality flaws (i.e., 1-P(NON)=51.6%). We observe that the most common reasons are image blur (i.e., 41.0%) and inadequate framing (i.e., 55.6%). In contrast, only a small portion of the images are labeled as too bright (i,e., 5.3%), too dark (5.6%), having objects obscuring the scene (3.6%), needing to be rotated for successful viewing (17.5%), or other reasons (0.8%). The statistics reveal the most promising directions for how to improve assistive photography tools to improve blind users' experiences. Specifically, the main functions should be focused on camera shake detection and object detection to mitigate the possibility of taking images with blur or framing flaws.

We also observe that the image quality issues are so severe that image content is deemed unrecognizable for 14.8% of the images. In absolute terms, this means that $3,829 and 379 hours of human annotation were wasted on employing crowdworkers to caption images that contained unrecognizable content.[1] In other words, great savings can be achieved by automatically filtering such uncaptionable images such that they are not sent to crowdworkers. We explore this idea further in Section 4.3.

Likelihood Image Has Unrecognizable Content Given its Quality-Flaw. We next examine the probability that an image's content is unrecognizable conditioned on each of the reasons for quality flaws. Results are shown in Figure 2.

Almost all reasons led to percentages that are larger than the overall percentage of unrecognizable images, which is 14.8% of all images. This demonstrates what we intuitively suspected, which is that images with quality flaws are more likely to have unrecognizable content. We observe that this trend is the strongest for images that suffer from obstructions (OBS) and inadequate lighting (BRT and DRK), with percentages just over 40%.

Interestingly, two categories have percentages that are smaller than the overall percentage of unrecognizable images, at 14.8% of all images. First, images that are flagged as needing to be rotated for proper viewing (ROT) have only 8.3% deemed unrecognizable. In retrospect, this seems understandable, as the content of images with a rotation flaw

---

[1]Crowdworkers were paid $0.132 for each image and spent an average of 47 seconds captioning each image.

could still be recognized if viewers tilt their heads (or apply visual display tools to rotate the images). Second, images labeled with no flaws (NON) have only 3.9% deemed unrecognizable. This tiny amount aligns with the concept that "unrecognizable" and "no flaws" are two conflicting ideas. Still, the fact the percentage is not 0% highlights that humans can offer different perspectives. Put differently, the image quality assessment task can be subjective.

Likelihood Image Has Each Quality-Flaw Given its Content is Unrecognizable. We next examine the probability that an image manifests each quality flaw given that its content is unrecognizable. Results are shown in Figure 2. Overall, our findings parallel those identified in the "Prevalence of Quality Issues" paragraph. For example, we
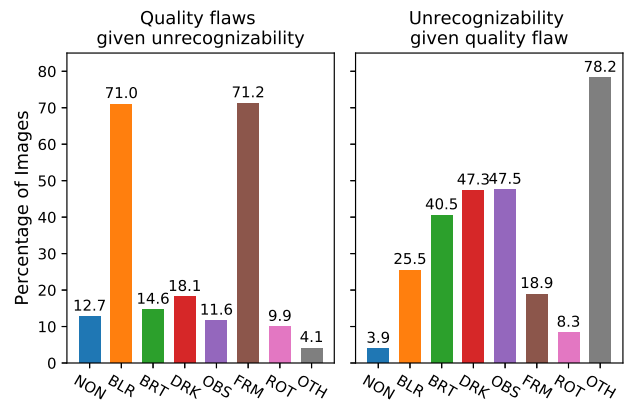

Figure 2: Left: Percentage of images with quality flaws given unrecognizability. Right: Percentage of unrecognizable images given quality flaws.
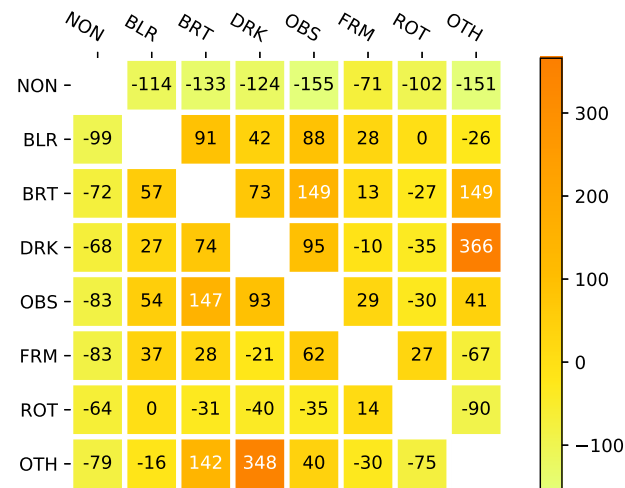

Figure 3: Interrelation of quality flaws. Values are scaled, with each multiplied by 100. The grid at the i-th row and the j-th column shows the value of $I($flaw i, flaw $j)$. The diagonal is suppressed for clarity.
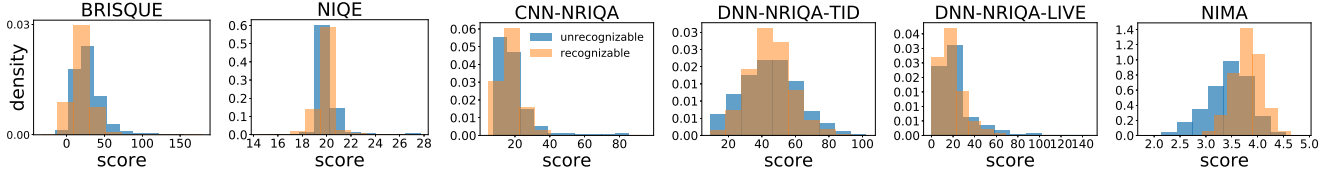
Figure 4: Distributions of image quality scores predicted by conventional NR-IQA systems [33, 34, 22, 6, 44] in our new VizWiz-QualityIssues dataset. The heavy overlap of the distributions of scores for recognizable and unrecognizable images reveals that none of the methods are able to distinguish recognizable images from unrecognizable images.

again observe the most common reasons are blurry images (71.0%) and improper framing (71.2%). Similarly, unrecognizable images are found to be associated less frequently with the other quality flaws.

Relationship Between Quality Flaws in Images. Finally, we quantify the relationship between all possible pairs of quality flaws. In doing so, we were motivated to provide a measure that offers insight into causality and co-occurrence when comparing any pair of quality flaws, while avoiding measuring joint probabilities. To meet this aim, we introduce a new measure which we call interrelation index $I(A, B)$, which is defined as follows:

$$I(A, B) = \frac{P(B|A)}{P(B)} - \frac{P(B|\bar{A})}{P(B)}. \quad (1)$$

More details about this measure and the motivation for it are provided in the Supplementary Materials. Briefly, larger positive $I(A, B)$ values indicate that A and B tend to co-occur with A causing B to happen more often. Results are shown in Figure 3.

We observe that almost all quality flaws tend to occur with one another, as shown with the positive values of I. At first, we were surprised to observe that there is a relationship between BRT and DRK (i.e., $I(\text{BRT}, \text{DRK}) = 73$ is greater than zero), since these flaws are seemingly incompatible concepts. However, from visual inspection of the data, we found some images indeed suffered from both lighting flaws. We exemplify this and other quality flaw correlations in the Supplementary Materials. From our findings, we also observe that "no flaws" does not co-occur with other quality flaws; i.e., the values in the grid are all negative for the row and column for NON. This finding aligns with our intuition that an image labeled with NON is less likely to have a quality flaw at the same time.

4. Classifying Unrecognizable Images

A widespread assumption when captioning images is that the image quality is good enough to recognize the image content. Yet, people who are blind cannot verify the quality of the images they take and it is known their images can be very poor in quality [5, 7, 17]. Accordingly, we

now examine the benefit of our large-scale quality dataset for training algorithms to detect when images are unrecognizable and so not captionable.

4.1. Motivation: Inadequate Existing Methods

Before exploring novel algorithms, it is important to first check whether existing methods are suitable for our purposes. Accordingly, we check whether related NR-IQA systems can detect when images are unrecognizable. To do so, we apply five NR-IQA methods on the complete VizWiz-QualityIssues dataset: BRISQUE [33], NIQE [34], CNN-NRIQA [22], DNN-NRIQA [6], and NIMA [44]. The first two are popular conventional methods that rely on hand-crafted features. The last three are based on neural networks and trained on IQA datasets mentioned in Section 2. For example, DNN-NRIQA-TID and DNN-NRIQA-LIVE in Figure 4 are trained on the TID dataset and LIVE dataset, respectively. Intuitively, if the algorithms are effective for this task, we would expect that the scores for recognizable images are distributed mostly in the high-score region, while the scores for unrecognizable images are distributed mostly in the low-score region.

Results are shown in Figure 4. A key finding is that the distributions of scores for recognizable and unrecognizable images heavily overlap. That is, none of the methods can distinguish recognizable images from unrecognizable images in our dataset. This finding shows that existing methods trained on existing datasets (i.e., LIVE, TID, CSIQ) are unsuitable for our novel task on the VizWiz-QualityIssues dataset. This is possibly in part because quality issues resulting from artificial distortions, such as compression, Gaussian blur, and additive Gaussian noise, differ from natural distortions triggered by poor camera focus, lighting, framing, etc. This also may be because there is no 1-1 mapping between scores indicating overall image quality and our proposed task, since an image with a low quality score may still have recognizable content.

4.2. Proposed Algorithm

Having observed that existing IQA methods are inadequate for our problem, we now introduce models for our novel task of assessing whether an image is recognizable.

Architecture. We use ResNet-152 [18] to extract image features, which are then processed by 2-dimensional global-pooling followed by two fully connected layers. The final layer is a single neuron with a sigmoid activation function.[2] We train this algorithm using an Adam optimizer with the learning rate set to 0.001 for 8 epochs. We fix the ResNet weights pre-trained on ImageNet [9] and only learn the weights in the two fully connected layers.

Dataset Splits. For training and evaluation of our algorithm, we apply a 52.5%/37.5%/10% split to our dataset to create the training, validation, and test splits.

Baselines. We compare our algorithm to numerous baselines. Included is random guessing, which means an image is unrecognizable with probability 0.148. We also analyze a linear SVM that predicts with scale-invariant feature transform (SIFT) features. Intuitively, a low-quality image should have few/no key points. We also evaluate a linear SVM that predicts from histogram of oriented gradients (HOG) features.

Evaluation Metrics. We evaluate each method using average precision, recall, and f1 scores. Accuracy is excluded because the distributions of unrecognizability are highly biased to "false" and such unbalanced data suffer from the accuracy paradox.

Results. Results are shown in Table 1. We observe that both SIFT and HOG are much stronger baselines than random guessing and get high scores on precision, especially 87.2 for SIFT. However, they both get low scores on recall. This means that SIFT and HOG are good at capturing a subset of unrecognizable images but still miss many others. On the other hand, the ResNet model gets much higher recall scores while maintaining decent average precision scores, implying that it is more effective at learning the characteristics of unrecognizable images.[3] This is exciting since such

---

[2]Due to space constraints, we demonstrate the effectiveness of this architecture for assessing the quality flaws in the Supplementary Materials. The primary difference for that architecture is that we replace ResNet-152 with XceptionNet [8], use three fully connected layers, and a final layer of eight neurons with eight sigmoid functions.

[3]Again, due to space constraints, results showing prediction performance for quality flaw classification is in the Supplementary Materials.

	Avg. precision	Recall	F1
ResNet-152	80.0	**75.1**	**71.2**
Random guessing	16.6	14.6	15.5
SIFT	**87.2**	42.3	56.9
HOG + linear SVM	56.4	41.2	47.6

Table 1: Performance of algorithms in assessing whether image content can be recognized (and so captioned).

an algorithm can be of immediate use to blind photographers, who otherwise must wait nearly two minutes to learn their image is unsuitable quality for image captioning.

4.3. Application: Efficient Dataset Creation

We now examine another potential benefit of our algorithm in helping to create a large scale training dataset.

To support this effort, we divide the dataset into three sets. One set is used to train our image unrecognizability algorithm. A second set is used to train our image captioning algorithms, which we call the captioning-training-set. The third set is used to evaluate our image captioning algorithms, which we call the captioning-evaluation-set.

We use our method to identify which images in the captioning-training-set to use for training image captioning algorithms. In particular, the N images flagged as recognizable are included and the remaining images are excluded. We compare this method to three baselines, specifically training on: all images in the captioning-training-set, a random sample of N images in the captioning-training-set, a perfect sample of N images in the captioning-training-set that are known to be recognizable images.

We evaluate two state-of-art image captioning algorithms, trained independently on each training set, with respect to eight evaluation metrics: BLEU-1-4 [35], METEOR [10], ROUGE-L [27], CIDEr-D [46], and SPICE [2].

Results are shown in Table 2. Our method performs comparably to when the algorithms were trained on all images as well as the perfect set. In contrast, our method yields improved results over the random sample. Altogether, these findings offer promising evidence that our prediction system is successfully retaining meaningful images while removing images that are not informative for the captioning task (i.e., unrecognizable). This reveals that a benefit of using the recognizability prediction system is to save time and money when crowdsourcing captions (by first removing unrecognizable images), without diminishing the performance of downstream trained image captioning algorithms.

5. Recognizing Unanswerable Visual Questions

The visual question "answerability" problem is to decide whether a visual question can be answered [17]. Yet, as exemplified in Figure 5, visual questions can be unanswerable because the image is unrecognizable or because the answer to the question is missing in a recognizable image. Towards enabling more fine-grained guidance to photographers regarding how to modify the visual question so it is answerable, we move beyond predicting *whether* a visual question is unanswerable [17] and introduce a novel problem of predicting *why* a visual question is unanswerable.

		B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr-D	SPICE
AoANet [19]	full training set	63.3	44.3	29.9	19.7	18.0	44.4	43.6	11.2
	perfect flag	63.3	43.8	29.5	19.9	18.1	44.2	43.6	11.5
	predicted flag	63.2	44.0	29.5	19.8	18.1	44.2	42.9	11.5
	random sample	62.5	43.3	28.8	18.9	18.0	44.1	41.9	11.4
SGAE [54]	full training set	62.8	43.3	28.6	18.8	17.3	44.0	32.4	10.4
	perfect flag	63.0	43.1	28.6	18.9	17.2	43.9	32.5	10.3
	predicted flag	63.1	43.1	28.4	18.7	17.2	44.0	32.4	10.4
	random sample	62.4	42.7	27.9	18.2	17.1	43.7	30.4	10.4

Table 2: Performance of two image captioning algorithms with respect to eight metrics trained on the full captioning-training-set, training images annotated to be recognizable (perfect flag), training images predicted to be recognizable (predicted flag), and a subset random sampled from the captioning-training-set. (B@ = BLEU-)
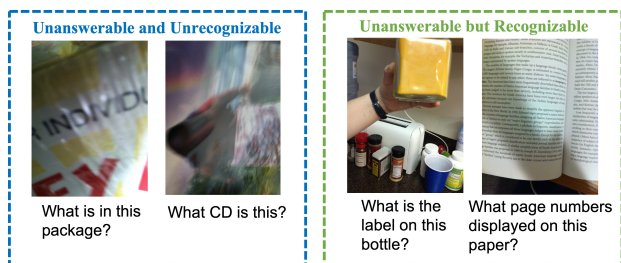

Figure 5: Examples of visual questions that are unanswerable for two reasons. The left two examples have unrecognizable images while the right two examples have recognizable images but the content of interest is missing from the field of view. Our posed algorithm correctly predicts why visual questions are unanswerable for these examples.

5.1. Motivation

We extend the VizWiz-VQA dataset [17], which labels each image-question pair as answerable or unanswerable. We inspect how answerability relates to recognizability and each quality flaw. For convenience, we use the following notations: A: answerable, $\bar{A}$: unanswerable, R: recognizable: $\bar{R}$: unrecognizable, Q: quality issues, and $P(\cdot)$: probability function. Results are shown in Figure 6. We can observe that for most quality flaws Q, $P(\bar{A}|Q)$ is larger than $P(\bar{A})$, and $P(\bar{A}) = 28.7\%$ increases to $P(\bar{A}|\bar{R}) = 58.7\%$. Additionally, the probability $P(\bar{R})$ increases from 14.8% to $P(\bar{R}|\bar{A}) = 30.2\%$ when questions are known to be unanswerable. Observing that a large reason for unanswerable questions is that images are unrecognizable images, we are motivated to equip VQA systems with a function that is able to clarify why their questions are unanswerable.

5.2. Proposed Algorithm

Algorithm. Our algorithm extends the Up-Down VQA model [3]. It takes as input encoded image features and a
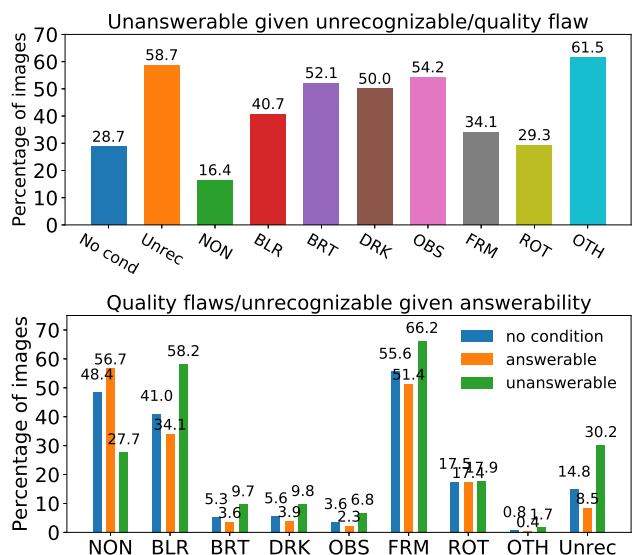

Figure 6: Top: Fractions of unanswerable questions conditioned on unrecognizability or a quality flaw. Bottom: Fractions of quality issues and unrecognizable images given answerability. Values are scaled by being multiplied with 100.

paired question. Image features could be grid-level features extracted by ResNet-152 [18] as well as object-level features extracted by Faster-RCNN [40] or Detectron [13, 52]. The input question is first encoded by a GRU cell. Then, a top-down attention module computes a weighted image feature from the encoded question representation and the input image features. The image and question features are coupled by element-wise multiplication. This coupled feature is processed by the prediction module to predict answerability and recognizability. We employ two different activation functions at the end of the model to make the final prediction. The first one is softmax which predicts three exclusive classes: answerable, unrecognizable, and insufficient content information (answers cannot be found in images). The

	Unans			Unrec given unans		
	AP	Rec	F1	AP	Rec	F1
[17]	71.7	–	64.8	–	–	–
Rand guess	–	–	–	31.1*	14.8	20.0
SIFT	–	–	–	**94.9***	45.3	61.3
HOG	–	–	–	73.1*	44.9	55.7
TD+soft	72.6	77.3	67.0	82.2	**79.3**	75.0
TD+sigm	73.6	71.2	**68.0**	86.6	**79.3**	78.6
BU+sigm	73.0	66.6	66.7	87.4	73.7	78.7
TD+BU+sigm	**74.0**	**82.3**	67.9	87.7	**79.3**	79.7
sigm w/o att.	67.7	66.1	64.2	86.7	66.7	74.2

TD: top-down attention. **BU**: bottom-up attention. **soft**: softmax. **sigm**: sigmoid. **att**: attention. **AP**: average precision. **Rec**: recall. **Unrec**: Unrecognizable. **Unans**: Unanswerable.

*: Precision is calculated, since true or false is predicted instead of a probability.

Table 3: Performance of predicting why a visual question is unanswerable: unrecognizable image versus unanswerable because the content of interest is missing from the field of view. [17] only predicts answerability and serves as the baseline for unanswerability prediction. Random guessing, SIFT, and HOG only predict recognizability and serve as the baselines for unrecognizability prediction.

second activation function is two independent sigmoids, one for answerability and the other for recognizability. We train the network using an Adam optimizer with a learning rate of 0.001, only for the layers after feature extraction.

Dataset Splits. We split VizWiz dataset into training/validation/test sets according to a 70%/20%/10% ratio.

Evaluation Metrics. We evaluate performance using average precision, precision, recall, and f1 scores, for which a simple threshold 0.5 is used to binarize probability values. For inter-model comparisons, we also report the precision-recall curve for each variant.

Baselines. For comparison, we consider a number of baselines. One approach is the original model for predicting whether a visual question is answerable, and also employs a top-down attention model [17]. We also evaluate the random guessing, SIFT, and HOG baselines used to evaluate the recognizability algorithms in the previous section.

Results. Results are shown in Table 3 and Figure 7. Our models perform comparably to the answerability baseline [17]. This is exciting because it shows that jointly learning to predict answerability with recognizability does not degrade the performance; i.e., the average precision scores from TD+softmax and TD+sigmoid models are better than
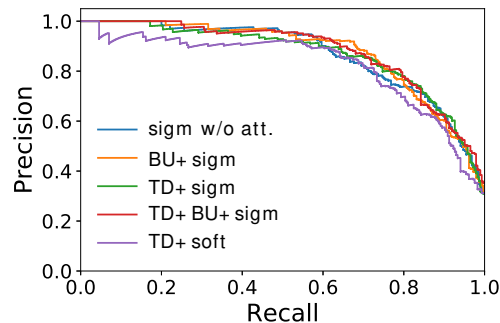

Figure 7: Precision-recall curves for five algorithms predicting unrecognizability when questions are unanswerable.

the one from the baseline [17] (72.6, 73.6 > 71.7) as well as the F1 scores (67.0, 68.0 > 64.8).

Our results also highlight the importance of learning to predict jointly the answerability with recognizability task (i.e., rows 5–9) over relying on more basic baselines (i.e., rows 2–4). As shown in Table 3, low recall values imply that SIFT and HOG fail to capture many unrecognizable images, while our models learn image features and excel in recall and f1 scores.

Next, we compare the results from TD+softmax and TD+sigmoid. We observe they are comparable in unanswerability prediction due to comparable average precision scores and F1 scores. For unrecognizability prediction, TD+softmax is a bit weaker than TD+sigmoid because due to slightly lower average precision and F1 scores. One reason for this may be the manual assignment of unrecognizability to false when answerability is true. Originally, 14.8% of images are unrecognizable, but after assignment, the portion drops to 8.7%. Learning from more highly biased data is a harder task, which could in part explain the weaker performance of TD+softmax model.

6. Conclusions

We introduce a new image quality assessment dataset that emerges from an authentic use case where people who are blind struggle to capture high-quality images towards learning about their visual surroundings. We demonstrate the potential of this dataset to encourage the development of new algorithms that can support real users trying to obtain image captions and answers to their visual questions. The dataset and all code are publicly available at `https://vizwiz.org`.

References

[1] http://www.taptapseeapp.com/. 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 7

[4] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280, 2019. 3

[5] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 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, 5

[6] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017. 2, 5

[7] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2117–2126. ACM, 2013. 1, 3, 5

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3, 6

[10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 6

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1, 3

[12] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 1, 2, 3

[13] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 7

[14] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 1, 3

[15] Danna Gurari and Kristen 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, 2017. 3

[16] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019. 3

[17] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey 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. 1, 3, 5, 6, 7, 8

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[19] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019. 7

[20] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 203–210. ACM, 2011. 3

[21] Dinesh Jayaraman, Anish Mittal, Anush K Moorthy, and Alan C Bovik. Objective quality assessment of multiply distorted images. In *2012 Conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 1693–1697. IEEE, 2012. 1, 2

[22] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 2, 5

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1, 3

[24] D Kundu, D Ghadiyaram, AC Bovik, and BL Evans. Large-scale crowdsourced study for high dynamic range images. *IEEE Trans. Image Process.*, 26(10):4725–4740, 2017. 2

[25] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. 1, 2

[26] Qiang Li and Zhou Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE journal of selected topics in signal processing*, 3(2):202–211, 2009. 2

[27] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014. 1, 3

[29] Lixiong Liu, Bao Liu, Hua Huang, and Alan Conrad Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014. 2

[30] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 1

[31] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016. 1, 2

[32] Lin Ma, Songnan Li, Fan Zhang, and King Ngi Ngan. Reduced-reference image quality assessment using reorganized dct-based image representation. *IEEE Transactions on Multimedia*, 13(4):824–829, 2011. 2

[33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 2, 5

[34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 2, 5

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 6

[36] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 1, 2

[37] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 1, 2

[38] Abdul Rehman and Zhou Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on Image Processing*, 21(8):3378–3389, 2012. 2

[39] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018. 2

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 7

[41] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality

assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 1, 2

[42] Rajiv Soundararajan and Alan C Bovik. Rred indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, 2011. 2

[43] Sundaram Suresh, R Venkatesh Babu, and Hyoung J Kim. No-reference image quality assessment using modified extreme learning machine classifier. *Applied Soft Computing*, 9(2):541–552, 2009. 2

[44] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 2, 5

[45] Marynel Vázquez and Aaron 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. 3

[46] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 6

[47] Zhou Wang and Alan C Bovik. Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40, 2011. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 2

[49] Zhou Wang and Eero P Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X*, volume 5666, pages 149–159. International Society for Optics and Photonics, 2005. 2

[50] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 2

[51] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Reduced-reference image quality assessment with visual information fidelity. *IEEE Transactions on Multimedia*, 15(7):1700–1705, 2013. 2

[52] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[53] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 1, 3

[54] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 7

[55] Peng Ye and David Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129–3138, 2012. 2

[56] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 2

[57] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2

[58] Yu Zhong, Pierre J Garrigues, and Jeffrey 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. 3

[59] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 3

[60] Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2110–2118, 2016. 1