Captioning Images Taken by People Who Are Blind

Danna Gurari, Yinan Zhao, Meng Zhang, Nilavra Bhattacharya

University of Texas at Austin

Abstract. While an important problem in the vision community is to design algorithms that can automatically caption images, few publicly-available datasets for algorithm development directly address the interests of real users. Observing that people who are blind have relied on (human-based) image captioning services to learn about images they take for nearly a decade, we introduce the first image captioning dataset to represent this real use case. This new dataset, which we call VizWiz-Captions, consists of over 39,000 images originating from people who are blind that are each paired with five captions. We analyze this dataset to (1) characterize the typical captions, (2) characterize the diversity of content found in the images, and (3) compare its content to that found in eight popular vision datasets. We also analyze modern image captioning algorithms to identify what makes this new dataset challenging for the vision community. We publicly-share the dataset with captioning challenge instructions at `https://vizwiz.org`.

1 Introduction

A popular computer vision goal is to create algorithms that can replicate a human's ability to caption any image [9, 29, 48]. Presently, we are witnessing an exciting transition where this dream of automated captioning is advancing into a reality, with automated image captioning now a feature available in several popular technology services. For example, companies such as Facebook and Microsoft are providing automated captioning in their social media [4] and productivity (e.g., Power Point) [1] applications to enable people who are blind to make some sense of images they encounter in these digital environments.

While much of the progress has been fueled by the recent creation of large-scale, publicly-available datasets (needed to train and evaluate algorithms), a limitation is that most existing datasets were created in contrived settings. Typically, crowdsourced workers were employed to produce captions for images curated from online, public image databases such as Flickr [6, 15, 21, 26, 28, 33, 56, 57]. Yet, we have observed over the past decade that people have been collecting image captions to meet their real needs. Specifically, people who are blind have sought descriptions[1] from human-powered services [2, 11, 45, 51, 58] to learn more about pictures they take of their visual surroundings. Unfortunately, images taken by these real users in the wild often exhibit dramatically different

[1] Throughout, we use "caption" and "description" interchangeably.

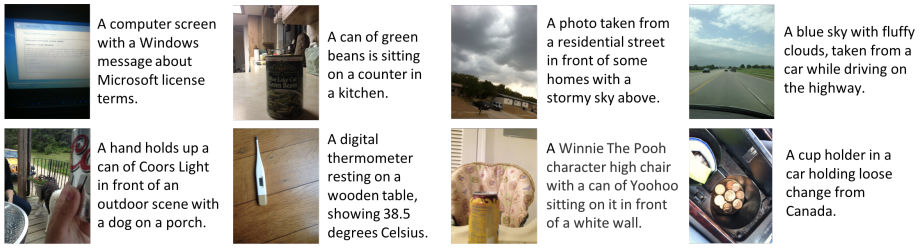

Fig. 1: Examples of captioned images in our new dataset, which we call VizWiz-Captions. These exemplify that images often contain text, exhibit a high variability in image quality, and contain a large diversity of content.

conditions than observed in the contrived environments used to design modern algorithms, as we will expand upon in this paper. Examples of some of the unique characteristics of images taken by real users of image captioning services are exemplified in Figure 1. The consequence is that algorithms tend to perform poorly when deployed on their images.

To address the above problem, we introduce the first publicly-available captioning dataset that consists of images taken by people who are blind. This dataset builds off of prior work which supported real users of a mobile phone application to submit a picture and, optionally, record a spoken question in order to learn about their images [11]. We crowdsourced captions for 39,181 images that were submitted. We also collected metadata for each image that indicates whether text is present and the severity of image quality issues to enable a systematic analysis around these factors. We call this dataset VizWiz-Captions.

We then characterize how our new dataset relates to the momentum of the broader vision community. To do so, we characterize how the captioned content relates/differs to what is contained in eight popular vision datasets that support the image captioning, visual question answering, and image classification tasks. We observe both that VizWiz-Captions shows many distinct visual concepts from those in existing datasets and regularly provides the answers to people's visual questions (Section 3.2). We also benchmark modern captioning algorithms, and find that they struggle to caption lower quality images.

We offer this work as a valuable foundation for designing more generalized computer vision algorithms that meet the large diversity of needs for real end users. Our dataset can facilitate and motivate progress for a broader number of scenarios that face similar complexities. For example, wearable lifelogging devices, autonomous vehicles, and robots also can result in varying image quality and many images showing textual information (e.g., street signs, billboards) as important real-world challenges that must be handled to solve downstream tasks.

To facilitate and encourage progress, we organized a dataset challenge and associated workshop to track progress and stimulate discussion about current research and application issues. Details about the dataset, challenge, and workshop can be found at the following link: `https://vizwiz.org`.

2 Related Work

Captioning Images for People Who are Blind. Given the clear wish from people who are blind to receive descriptions of images [3, 10, 11, 13, 37, 42, 52], many human-in-the-loop [2, 3, 5, 38, 45, 51] and automated services [1, 4] have emerged to do so. A challenge shared across such services is what content to describe. Although there remains a lack of guidance for images taken by people who are blind [49], it is known that many people who are blind report a preference to receive descriptions of images over nothing (even if inaccurate) [23, 45, 46, 53]. Accordingly, to facilitate progress on automated solutions for captioning images taken by this population, we introduce a new dataset to represent this use case. In doing so, we aim to support the design of a cheaper, faster, and more private alternative than is possible with human-based captioning services.

Image Captioning Datasets. Over the past decade, nearly 20 publicly-shared captioning datasets have been created to support the development of automated captioning algorithms [6, 14, 15, 18, 19, 21, 26–28, 32, 33, 43, 47, 56, 57, 60]. The trend has been to include a larger number of examples, relying on scraping images from the web (typically Flickr) to support the growth from a few thousand [19, 20, 43] to hundreds of thousands [15, 27, 33] of captioned images in such datasets. In doing so, such work has strayed from focusing on real use cases. To help align the vision community to focus on addressing the real interests of people who need image captions, we instead focus on introducing a captioning dataset that emerges from a natural use case.

Accordingly, our work more closely aligns with the earlier datasets that emerged from authentic image captioning scenarios. This includes captioned images in newspaper articles [20] and provided by tour guides about photographs of tourist locations [22]. Unlike these prior works, we focus on a distinct use case (i.e., captioning blind photographers' images) and our new dataset is considerably larger (i.e., contains nearly 40,000 images versus 3,361 [20] and 20,000 [22]).

More generally, to our knowledge, our new captioning dataset is the first that comes with metadata indicating for each image whether text is present and the severity of image quality issues, thereby enabling systematic analysis around these factors. We expect this new dataset will contribute to the design of more robust, general-purpose captioning algorithms.

Content in Vision Datasets. The typical trend for curating images for popular vision datasets is to scrape various web search engines for pre-defined categories/search terms. For example, this is how popular object recognition datasets (e.g., ImageNet [44] and COCO [36]), scene recognition datasets (e.g., SUN [54] and Places205 [59]), and attribute recognition datasets (e.g., SUN-attributes [41] and COCO-attributes [40]) were created. Observing that automated methods rely on such large-scale datasets to guide what concepts they learn, a question emerges of how well the content in such contrived datasets reflect the interests of real users of image descriptions services. We conduct comparisons between popular vision datasets and our new dataset to provide such insight. This analysis is valuable both for highlighting the value of existing datasets to support a real use case and revealing how vision datasets can be improved.

3 VizWiz-Captions

We now introduce VizWiz-Captions, a dataset that consists of descriptions about images taken by people who are blind. Our work builds upon two existing datasets that contain images taken by real users of a visual description service [24, 25]. The images in these datasets originate from users of the mobile phone application VizWiz [11], who each submitted a picture with, optionally, a recorded spoken question in order to receive a description of the image or answer to the question (when one was asked) from remote humans. In total, we used the 39,181 images that are publicly-shared and were not corrupted to obfuscate private content. Of these, 16% (i.e., 6,339) lack a question. We detail below our creation and analysis of this dataset.

3.1 Dataset Creation

Image Captioning System. To collect captions, we designed our captioning task for use in the crowdsourcing platform Amazon Mechanical Turk (AMT). To our knowledge, the only public precedent for crowdsourcing image descriptions from crowdworkers for images taken by people who are blind is the VizWiz mobile phone application [11]. This system offered vague instructions to 'describe the image'. Given this vague precedence, we chose to adapt the more concrete task design from the vision community, as described below.

We employed the basic task interface design used by prior work in the vision community [6, 15, 28, 57]. It displays the image on the left, instructions on the right, and a text entry box below the instructions for entering the description. The instructions specify to include at least eight words as well as what not to do when creating the caption (e.g., do not speculate what people in the image might be saying/thinking or what may have happened in the future/past).

We further augmented the task interface to tailor it to unique characteristics of our captioning problem. These augmentations resulted both from consultation with accessibility experts and iterative refinement over four pilot studies. First, to encourage crowdworkers to address the interests of the target audience, we added the instruction to "Describe all parts of the image that may be important to a person who is blind." Second, to encourage crowdworkers to focus on the content the photographer likely was trying to capture rather than any symptoms of low quality images that inadvertently arise for blind photographers, we instructed crowdworkers "DO NOT describe the image quality issues." However, given that some images could be insufficient quality for captioning, we provided a button that the crowdworker could click in order to populate the description with pre-canned text that indicates this occurred (i.e., "Quality issues are too severe to recognize visual content."). Next, to discourage crowdworkers from performing the optical character recognition problem when text is present, we added the following instruction: "If text is in the image, and is important, then you can summarize what it says. DO NOT use all the specific phrases that you see in the image as your description of the image." Finally, to enrich our analysis, we

asked crowdworkers to provide extra information about each image regarding whether text is present.

Caption Collection and Post-Processing. For each of the 39,181 images, we collected redundant results from the crowd. In particular, we employed five AMT crowdworkers to complete our task for every image. We applied a number of quality control methods to mitigate concerns about the quality of the crowdsourced results, summarized in the Supplementary Materials. In total, we collected 195,905 captions. All this work was completed by 1,623 crowdworkers who contributed a total of 3,736 person-hours. With it being completed over a duration of 101.52 hours, this translates to roughly 37 person-hours of work completed every hour. We post-processed each caption by applying a spell-checker to detect and fix misspelled words.

3.2 Dataset Analysis

Quality of Images. We first examined the extent to which the images were deemed to be insufficient quality to caption. This is important to check, since people who are blind cannot verify the quality of the images they take, and it is known their images can be poor quality due to improper lighting (i.e., mostly white or mostly black), focus, and more [12, 16, 25]. To do so, we tallied how many of the five crowdworkers captioned each image with the pre-canned text indicating insufficient quality for captioning (i.e., "Quality issues are too severe..."). The distribution of images for which none to all 5 crowdworkers used this pre-canned text is as follows: 68.5% for none, 16.7% for 1 person, 5.9% for 2 people, 3.6% for 3 people, 3.1% for 4 people, and 2.2% for all 5 people.

We found that the vast majority of images taken by blind photographers were deemed good enough quality that the content can be recognized. Only 9% of the images were deemed insufficient quality for captioning by the majority of the crowdworkers. A further 22.6% of images were deemed insufficient quality by a minority of the crowdworkers (i.e., 1 or 2). Altogether, these findings highlight a range of difficulty for captioning, based on the extent to which crowdworkers agreed the images are (in)sufficient quality to generate a caption. In Section 4, we report the ease/difficulty for algorithms to caption images based on this range of perceived difficulty by humans.

VizWiz-Captions Characterization. Next, we characterized the caption content. For this purpose, we excluded from our analysis all captions that contain the pre-canned text about insufficient quality images ("Quality issues are too severe...") as well as those that were rejected. This resulted in a total of 168,826 captions.

We first quantified the composition of captions, by examining the typical description length as well as the typical number of objects, descriptors, actions, and relationships. To do so, we computed as a proxy the average number of words as well as the average number of nouns, adjectives, verbs, and spatial relation words per caption. Results are shown in Table 1 (row 1). Our findings reveal that sentences typically consist of roughly 13 words that involve four to five

	Average Count Per Image					Unique Count for All Images				
	words	nouns	verbs	adj	spa-rel	words	nouns	verbs	adj	spa-rel
Ours	13.0	4.4	0.9	1.4	1.9	24,422	16,400	4,040	8,755	275
Ours-WithQues	13.0	4.4	0.9	1.4	1.9	22,261	14,933	3,719	7,882	244
Ours-NoQues	13.0	4.4	0.9	1.5	1.9	10,651	7,249	1,616	3,212	120
Ours-WithText	12.9	4.5	0.9	1.4	1.9	21,161	14,277	3,294	7,263	243
Ours-NoText	13.1	4.2	0.9	1.6	1.9	10,711	7,114	1933	3,508	127
[15]-All	11.3	3.7	1.0	0.9	1.7	30,122	19,998	6,697	9,651	381
[15]-Sample	11.3	3.7	1.0	0.9	1.7	16,966	11,211	3,822	4,922	197

Table 1: Characterization of our VizWiz-Captions dataset. Shown is the average count per caption as well as the total count of unique words, nouns, verbs, adjectives, and spatial relation words for each dataset with respect to all captions, various subsets to support finer-grained analysis, and MSCOCO-Captions dataset [15] for comparison. (adj = adjectives; spa-rel = spatial relations)

objects (i.e., nouns) in conjunction with one to two descriptors (i.e., adjectives), one action (i.e., verb), and two relationships (i.e., spatial relationship words). Examples of sentences featuring similar compositions include "A hand holding a can of Ravioli over a counter with a glass on it" and "Red car parked next to a black colored SUV in an outside dirt parking lot."

We enriched our analysis by examining the typical caption composition separately for the 16% (i.e., 6,339) of images that originated from a captioning use case and the remaining 84% of images that originated from a VQA use case (meaning the image came paired with a question). Results are shown in Table 1, rows 2–3. We observe that the composition of sentences is almost identical for both use cases. This offers encouraging evidence that the images taken from a VQA setting are useful for large-scale captioning datasets.

We further enriched our analysis by examining how the caption composition changes based on whether the image contains text. We deemed an image as containing text if the majority of the five crowdworkers indicate it does. In our dataset, 63% (24,812) of the images contain text. The caption compositions for both subsets are shown in Table 1, rows 4–5. Our findings reveal that images containing text tend to have more nouns and fewer adjectives than images that lack text. Put differently, the presence of text appears to be more strongly correlated to the object recognition task. We hypothesize this is in part because crowdworkers commonly employ both a generic object recognition category followed by a specific object category gleaned from reading the text when creating their descriptions; e.g., "a box of Duracell procell batteries" and "a can of Ravioli." It's also possible that text is commonly present in more complex scenes that show a greater number of objects.

We also quantified the diversity of concepts in our dataset. To do so, we report parallel analysis to that above, with a focus on the *absolute number of unique* words, nouns, adjectives, verbs, and spatial relation words across all captions.

Results are shown in the right half of Table 1. These results demonstrate that the dataset captures a large diversity of concepts, with over 24,000 unique words. We visualize the most popular words in the Supplementary Materials, and conduct further analysis below to offer insight into how these concepts relate/differ to those found in popular computer vision datasets.

Comparison to Popular Captioning Dataset. We next compared our dataset to the popular MSCOCO-Captions dataset [15], and in particular the complete MSCOCO training set for which the captions are publicly-available.

Paralleling our analysis of VizWiz-Captions, we quantified the average as well as total unique number of words, nouns, adjectives, verbs, and spatial words in MSCOCO-Captions [15]. To enable side-by-side comparison, we not only analyzed the entire MSCOCO-Captions training set but also randomly sampled the same number of images with the identical distribution of number of captions per image as was analyzed for VizWiz-Captions. We call this subset MSCOCO-Sample. Results are shown in Table 1, rows 6–7. The results reveal that VizWiz-Captions tends to have a larger number of words per caption than MSCOCO-Captions; i.e., an average of 13 words versus 11.3 words. This is true both for the full set as well as the sample from MSCOCO-Captions. As shown in Table 1, the greater number of words is due to a greater number of nouns, adjectives, and spatial relation words per caption in VizWiz-Captions. Possible reasons for this include that the images show more complex scenes and that crowdworkers were motivated to provide more descriptive captions when knowing the target audience is people who are blind.

We additionally measured the content overlap between the two datasets. Specifically, we computed the percentage of words that appear both in the most common 3,000 words for VizWiz-Captions and the most common 3,000 words in MSCOCO-Captions. The overlap is 54.4%. This finding underscores a considerable domain shift in the content that blind photographers take pictures of and what artificially constructed datasets represent. We visualize examples of novel concepts not found in MSCOCO-Captions in the Supplementary Materials.

We also assessed the similarity of captions generated by different humans using the specificity score [31] for both our dataset and MSCOCO-Captions. Due to space constraints, we show the resulting distributions of scores in the Supplementary Materials for both datasets. In summary, the scores are similar.

Comparison to Visual Question Answering Dataset. Given that 84% of the images originate from a VQA use case (i.e., where a question was also submitted about the image), our new dataset offers a valuable test bed to explore the potential for generic image captions to answer users' real visual questions. Accordingly, we explore this for each image in our dataset for which we both have publicly-available answers for the question and the question is deemed to be "answerable" [24, 25].

We first evaluate this using a *quantitative measure*. Specifically, for the 24,842 answerable visual questions in the publicly-available training and validation splits, we tally the percentage for which the answer can be found in at least

	All Images				Images With Text				Images Without Text			
	All	Yes/No	#	Other	All	Yes/No	#	Other	All	Yes/No	#	Other
Quant	33%	1%	8%	34%	35%	1%	10%	36%	30%	1%	3%	31%
Qual	32%	35%	23%	38%	–	–	–	–	–	–	–	–

Table 2: Percentage of VQAs for which an image caption contains the answer with respect to both a quantitative ("Quant") and qualitative ("Qual") analysis. Fine-grained quantitative analysis is shown based on the type of answer that is elicited by the visual question (i.e., "yes/no", "#", and "other") as well as based on whether the images contain text. (# = number)

one of the five captions using exact string matching. We set the answer to the most popular answer from the 10 provided with each visual question. We conduct this analysis with respect to all images as well as separately for only those images which are paired with different answer types for the visual questions—i.e., "yes/no" (860 images), "number" (314 images), and "other" (23,668 images). Results are shown in Table 2. Overall, we observe that captions contain the information that people who are blind were seeking for roughly one third of their visual questions. This sets a lower bound, since string matching is an extremely rigid scheme for determining whether text matches.

We perform parallel quantitative analysis based on whether images contain text. For visual questions that contain text (i.e., 15,910 answerable visual questions), we again analyze the visual questions that lead to "yes/no" (447 images), "number" (218 images), and "other" (15,245 images) answers. We also perform this analysis on only the subset of visual questions that lack text (i.e., 8,932 answerable visual questions)—i.e., "yes/no" (413 images), "number" (96 images), and "other" (8,423 images). Results are shown in Table 2. We observe that the answer tends to be contained in the caption more often when the image contains text. This discrepancy is the largest for "number" questions, which we hypothesize is due to images showing currency. People seem to naturally want to characterize how much money is shown for such images, which conveniently is the information sought by those asking the questions.

To also capture when the answer to a visual question is provided implicitly in the captions, we next used a *qualitative approach*. We sampled 300 visual questions, with 100 for each of the three answer types.[2] Then, one of the authors reviewed each visual question with the answers and five captions to decide whether each visual question was answered by any of the captions about the image. Results are shown in Table 2, row 4. We observe a big jump in percentage for "yes/no" and "number" questions. The greatest boost is observed for "yes/no"

[2] For "yes/no" visual questions, we sampled 50 that have the answer "yes" and another 50 with the answer "no." For "number" visual questions, we sampled 50 that begin with the question "How many" and another 50 that begin with "How much." Finally, we randomly sampled another 100 visual questions from the "other" category.

visual questions where the percentage jumps from 0% to 35%. We attribute this to the "yes" questions more than the "no" questions–i.e., 22/50 for "yes" and 13/50 for "no"—since content that is asked about may be described when it is present in the image but will almost definitely not be described when it is not. Still, "no" questions often arise because, when the answer can be inferred, the caption typically also answers a valuable follow-up question. For example, a caption that states "A carton of banana flavored milk sits in a clear container with eggs" arguably answers the question "Is this chocolate milk?" (i.e., the answer is "no") while providing additional information (i.e., it is "banana milk").

Altogether, our findings show that at least one third of the visual questions can be answered with image captions. In other words, the captions regularly provide useful information for people who are blind. We attribute this large percentage partly to the fact that many questions for VQA just paraphrase a request to complete the image captioning task; e.g., nearly half of the questions ask a variant of "what is this" or "describe this" [25]. It also may often be obvious to the people providing captions what information the photographer was seeking when submitting the image with a question. Regardless of the reason though, it appears the extra work of devising a question regularly can be unnecessary in practice. A valuable direction for future research is to continue improving our understanding for how to align captions with real end users' interests.

Comparison to Popular Image Classification Datasets. Observing that automated captioning algorithms often build off of pretrained modules that perform more basic tasks such as image classification and object detection (e.g., trend dates back at least to Baby Talk [34] in 2013), we next examine the overlap between concepts in VizWiz-Captions and popular vision datasets that often are used to train such modules. For our analysis, we focus on three visual tasks: recognizing objects, scenes, and attributes.

We began by tallying how many popular concepts from existing vision datasets for the three vision tasks are found in VizWiz-Captions. To do so, we computed matches using extract string matching. When comparing concepts in VizWiz-Captions to the object categories that span both ImageNet [44] and COCO [36], we found that all nine categories that are shared across the two datasets are also found in VizWiz-Captions. Similarly, we found that all scene categories which span both SUN [54] and Places205 [59] (i.e., 70 categories) are captured in VizWiz-Captions. Additionally, all attribute categories that span both COCO-Attributes [41] and SUN-Attributes [40] (i.e., 14 categories) are captured in VizWiz-Captions. Consequently, across all three tasks, all concepts that are shared across the pair of mainstream vision datasets are also present in VizWiz-Captions. This is interesting in part because VizWiz-Captions was not created with any of these tasks in mind. Moreover, it underscores the promise for models trained on existing datasets to generalize well in recognizing some content encountered by people who are blind in their daily lives.

We also tally how many of the images in each dataset contain the popular concepts discussed above. Results are reported with respect to each of three

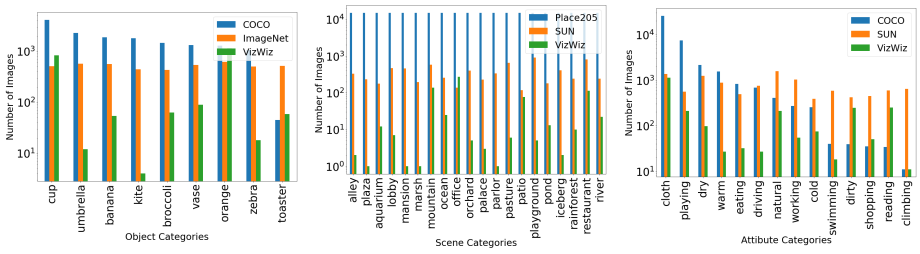

Fig. 2: Histogram showing how many images in our VizWiz-Captions as well as popular vision datasets contain each category for the following vision problems: (a) object recognition, (b) scene classification, and (c) attribute recognition.

classification tasks in Figure 2.[3] As shown, the number of examples in VizWiz-Captions is typically considerably fewer than observed for the other two popular datasets per task. This is not entirely surprising given that the absolute number of images in VizWiz-Captions is at least an order of magnitude smaller than most of the datasets (i.e., the object and scene classification datasets). We offer this analysis as a lower bound since *explicitly* asking crowdworkers whether each category is present could reveal a greater prevalence of these concepts. Observing that relying on data from real use cases alone likely provides an insufficient number of examples per category to successfully train algorithms, this finding highlights a potential benefit of contrived datasets in supplementing examples to our real-world dataset. We leave this idea as a valuable area for future work.

We also computed the percentage of all categories from each of the classification datasets that are captured by VizWiz-Captions. Again, we used exact string matching to do so. For object recognition, VizWiz-Captions contains only 1% of the categories in ImageNet and 11% of those in COCO. For scene recognition, VizWiz-Captions contains only 18% of the categories in SUN and 34% of those in Places205. For attribute recognition, VizWiz-Captions contains only 14% of the categories in COCO-Attributes and 7% of those in SUN-Attributes. Observing that these vision datasets are reserved to a range of hundreds to at most a thousand categories while we know from Table 1 that VizWiz-Captions contains thousands of unique nouns and adjectives, these datasets appear to provide very little coverage for the diversity of content captured in VizWiz-Captions. Altogether, these findings offer promising evidence that existing contrived image classification datasets provide a considerable mismatch to the concepts encountered by blind users who are trying to learn about their visual surroundings. Our findings serve as an important reminder that much progress is still needed to accommodate the diversity of content found in real-world settings.

[3] We show parallel analysis in the Supplementary Materials using the proportions of each dataset rather than absolute numbers. For both sets of results, we only show a subset of the 70 scene categories.

4 Algorithm Benchmarking

We next benchmarked state-of-art image captioning algorithms to gauge the difficulty of VizWiz-Captions and what makes it difficult for modern algorithms.

Dataset Splits. Using the same test set as prior work [25], we applied roughly a 70%/10%/20% split for the train/val/test sets respectively, resulting in a 23,431/7,750/8,000 split. To focus algorithms on learning novel captions, we exclude from training and evaluation captions with pre-canned text about insufficient quality images or rejected ones that were deemed spam.

Baselines. We benchmarked nine algorithms based on three modern image captioning algorithms that have been state-of-art methods for the MSCOCO-Captions [15] challenge: Up-Down [8], SGAE [55], and AoANet [30]. Up-Down [8] combines bottom-up and top-down attention mechanisms to consider attention at the level of objects and other salient image regions. SGAE [55] relies on a Scene Graph Auto-Encoder (SGAE) to incorporate language bias into an encoder-decoder framework, towards generating more human-like captions. AoANet [30] employs an Attention on Attention (AoA) module to determine the relevance between attention results and queries. We evaluated all three algorithms, which originally were trained on the MSCOCO-Captions dataset, *as is*. These results are useful in assessing the effectiveness of the MSCOCO training dataset for teaching computers to describe images taken by people who are blind. We also *fine-tuned* each pretrained network to VizWiz-Captions and *trained each network from scratch* on VizWiz-Captions. These algorithms are helpful for assessing the usefulness of each model architecture for describing images taken by people who are blind. For all algorithms, we used the publicly-shared code and default training hyper-parameters reported by the authors. We also benchmarked a commercial text detector[4] on test images containing text.

Evaluation. We evaluated each method with eight metrics that often are used for captioning: BLEU-1-4 [39], METEOR [17], ROUGE-L [35], CIDEr-D [50], and SPICE [7].

Overall Performance. We report the performance of each method in Table 3.

Observing the performance of existing algorithms that are *pretrained* on MSCOCO-Captions [15], we see that they can occasionally accurately predict captions for images coming from blind photographers. This is exciting as it shows that progress on artificially-constructed datasets can translate to successes in real use cases. We attribute the prediction successes to when the images are both good quality and show objects that are common in MSCOCO-Captions, as exemplified in the top six examples in Figure 3.

We consistently observe considerable performance improvements from the algorithms when training them on VizWiz-Captions, including when they are

[4] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

		B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr	SPICE
[8]	pretrained	52.8	32.8	19.2	11.3	12.6	35.8	18.9	5.8
	from scratch	64.1	44.6	30.0	19.8	18.4	43.2	49.7	12.2
	fine-tuned	62.1	42.3	28.2	18.6	18.0	42.0	48.2	11.6
[55]	pretrained	55.8	36.0	21.8	13.5	13.4	38.1	20.2	5.9
	from scratch	67.3	48.1	33.2	22.8	19.4	46.6	52.4	12.8
	fine-tuned	**68.5**	**49.4**	**34.5**	**23.9**	20.2	**47.3**	**61.2**	13.5
[30]	pretrained	54.9	34.7	21.0	13.2	13.4	37.6	19.4	6.2
	from scratch	66.4	47.9	33.4	23.2	**20.3**	47.1	60.5	**14.0**
	fine-tuned	66.6	47.4	32.9	22.8	19.9	46.6	57.6	13.7

Table 3: Performance of top-performing image captioning algorithms on the VizWiz-Captions test set with respect to eight metrics. Results are shown for three variants of the algorithms: when they are pre-trained on MSCOCO-Captions [15], trained only on the VizWiz-Captions dataset, and pre-trained on MSCOCO-Captions followed by fine-tuning to the VizWiz-Captions dataset. (B@ = BLEU-)

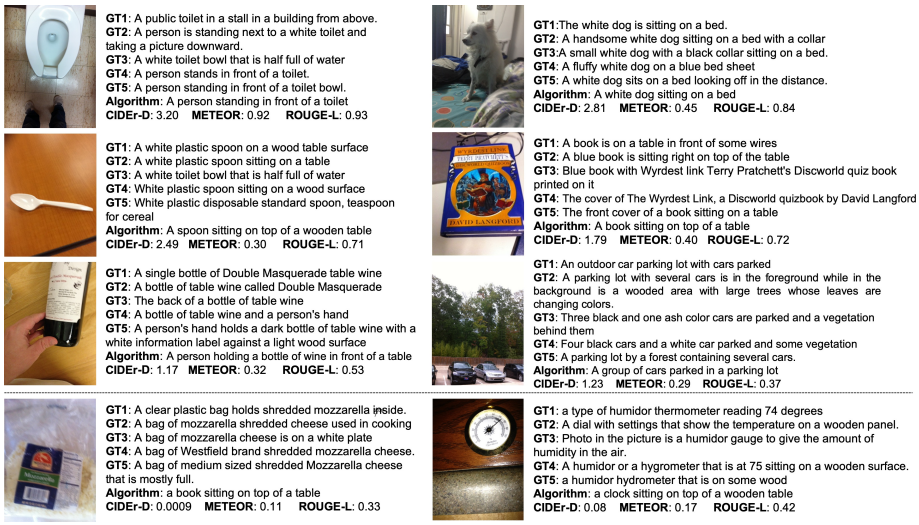

Fig. 3: Examples of a state-of-art image captioning algorithm's successes (top three rows) and failures (bottom row) in generating captions for images taken in a real use case. Results are for SGAE [55] pretrained on MSCOCO-Captions.

trained from scratch and fine-tuned. For instance, we observe roughly a 10 percentage point boost with respect to BLEU-1 and 30 percentage point boost with respect to CIDEr-D across the three algorithms. Still, the scores are considerably lower than what is observed when these same algorithmic frameworks are evaluated on the MSCOCO-Captions test set. For example, we observe the

BLEU-1 score is over 20 percentage points lower and the METEOR score is almost 20 percentage points lower, when comparing the performance of the top-performing algorithm for VizWiz-Captions against the top-performing algorithm for MSCOCO-Captions (i.e., AoANet [30]). This finding highlights that VizWiz-Captions currently offers a challenging dataset for the vision community.

When comparing outcomes between algorithms that are trained from scratch on VizWiz-Captions versus fine-tuned to VizWiz-Captions, we do not observe a considerable difference. For instance, we observe better performance when Up-Down [8] and AoANet [30] are trained from scratch on VizWiz-Captions rather than fine-tuned from models pretrained on MSCOCO-Captions, and vice versa for SGAE [55]. We found it surprising there is similar performance, given that VizWiz-Captions is roughly one order of magnitude smaller than MSCOCO-Captions. Valuable areas for future work include investigating the benefit of domain adaptation methods as well as how to successfully leverage larger contrived datasets (e.g., MSCOCO-Captions) to improve the performance of algorithms on VizWiz-Captions.

Fine-Grained Analysis. We enriched our analysis to better understand why algorithms struggle to accurately caption the images. To do so, we evaluated the top-performing algorithms for VizWiz-Captions with respect to two characteristics. First, we characterized performance independently for images in the test set based on whether they are flagged as containing text. We also characterized performance independently for images flagged as different difficulty levels, based on the number of crowdworkers who deemed the images insufficient quality to generate a meaningful caption; i.e., easy is when all five people generated novel captions, medium when 1-2 crowdworkers flagged the images as insufficient quality for captioning, and difficult when 3-4 crowdworkers flagged the images as insufficient quality. Results are shown in Table 4 for the top algorithms from Table 3; i.e., "from scratch" for [8] and [30] and "fine-tuned" for [55].

We observe two trends for the performance of algorithms based on whether text is present. We find the text detector does very poor, underscoring a key challenge for designing algorithms is to figure out how to integrate knowledge about text into captions. In contrast, we find that all captioning algorithms perform better when text is present. Initially, we found this surprising given that none of the benchmarked algorithms were designed to handle text (e.g., by incorporating an optical recognition module). Moreover, images with text cover many more unique concepts than images lacking text, as shown in Table 1. We hypothesize the improved performance is because images containing text provide a simpler domain that conforms to a fewer set of templates for the captions. For example, from visual inspection, we observe captions for such images often include "a box/bag of ... on/in ...". The captioning patterns for this simpler domain may be easier to learn for the algorithms. If so, this underscores an inadequacy of current evaluation metrics and a need for new metrics that prioritize the information people who are blind want.

When observing algorithm performance based on the captioning difficulty level, we find it parallels human difficulty with algorithms performing best on

		B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr	SPICE
[8]	WithText	65.7	46.6	32.3	21.7	19.2	45.1	49.3	12.6
	LackText	60.2	40.0	25.2	15.7	16.9	39.7	46.7	11.4
	Easy	67.8	48.3	33.2	22.2	19.5	45.7	53.2	12.5
	Medium	60.8	40.0	25.4	16.2	17.2	40.4	45.8	11.9
	Difficult	32.1	16.9	9.2	5.4	10.6	26.2	34.7	9.0
[55]	WithText	69.9	51.2	36.6	25.8	21.0	49.3	62.2	14.1
	LackText	65.7	45.9	30.3	20.2	18.7	43.7	55.5	12.5
	Easy	72.2	53.3	38.0	26.7	21.4	50.0	65.8	13.9
	Medium	65.1	44.7	29.3	19.3	18.8	44.3	55.8	13.3
	Difficult	35.6	20.1	11.7	7.4	11.9	28.8	42.7	9.3
[30]	WithText	68.3	50.1	35.8	25.3	21.3	49.2	62.7	14.6
	LackText	63.0	43.7	28.8	18.9	18.5	43.3	52.5	12.9
	Easy	69.8	51.4	36.5	25.6	21.4	49.6	64.3	14.3
	Medium	63.6	44.0	29.2	19.5	19.2	44.5	56.0	14.2
	Difficult	36.0	20.3	11.8	7.6	12.2	29.6	44.3	10.6
Text_API	WithText	14.9	8.8	5.7	3.9	10.4	15.9	24.6	–

Table 4: Analysis of the top-performing image captioning algorithms and a text detection algorithm based on whether images contain text and the image "difficulty". (B@ = BLEU-)

the easiest images for humans. While not surprising, this finding underscores the practical difficulty of designing algorithms that can handle low quality images, which we know are somewhat common from real users of image captioning services (i.e., people who are blind).

5 Conclusions

We offer VizWiz-Captions as a valuable foundation for designing image captioning algorithms to support a natural, socially-important use case. More broadly, our analysis reveals important problems that the vision community needs to address in order to deliver more generalized algorithms. Interesting future work includes holistically improving vision solutions to include consideration of potentially, valuable additional sensors to more effectively meet real users' needs (e.g., GPS, sound waves, infrared).

Acknowledgements. We thank Meredith Ringel Morris, Ed Cutrell, Neel Joshi, Besmira Nushi, and Kenneth R. Fleischmann for their valuable discussions about this work. We thank Peter Anderson and Harsh Agrawal for sharing their code for setting up the EvalAI evaluation server. We thank the anonymous crowdworkers for providing the annotations. This work is supported by National Science Foundation funding (IIS-1755593), gifts from Microsoft, and gifts from Amazon.

References

1. Add alternative text to a shape, picture, chart, SmartArt graphic, or other object. https://support.office.com/en-us/article/add-alternative-text-to-a-shape-picture-chart-smartart-graphic-or-other-object-44989b2a-903c-4d9a-b742-6a75b451c669.
2. BeSpecular. https://www.bespecular.com.
3. Home - Aira : Aira. https://aira.io/.
4. How does automatic alt text work on Facebook? — Facebook Help Center. https://www.facebook.com/help/216219865403298.
5. TapTapSee - Blind and Visually Impaired Assistive Technology - powered by the CloudSight.ai Image Recognition API. https://taptapseeapp.com/.
6. H. Agrawal, K. Desai, X. Chen, R. Jain, D. Batra, D. Parikh, S. Lee, and P. Anderson. Nocaps: Novel object captioning at scale. *arXiv preprint arXiv:1812.08658*, 2018.
7. P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.
8. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
9. S. Bai and S. An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.
10. C. L. Bennett, M. E. Mott, E. Cutrell, and M. R. Morris. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 76. ACM, 2018.
11. 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.
12. E. Brady, M. R. Morris, Y. Zhong, S. White, and J. 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.
13. 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.
14. J. Chen, P. Kuznetsova, D. Warren, and Y. Choi. Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, 2015.
15. 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.
16. T.-Y. Chiu, Y. Zhao, and D. Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656, 2020.
17. M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

18. D. Elliott and F. Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, 2013.
19. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer, 2010.
20. Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. *Proceedings of ACL-08: HLT*, pages 272–280, 2008.
21. C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.
22. M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Int. Workshop OntoImage*, volume 5, 2006.
23. D. Guinness, E. Cutrell, and M. R. Morris. Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 518. ACM, 2018.
24. D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. 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.
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.
26. D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE, 2015.
27. W. Havard, L. Besacier, and O. Rosec. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *arXiv preprint arXiv:1707.08435*, 2017.
28. M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
29. M. D. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):118, 2019.
30. L. Huang, W. Wang, J. Chen, and X.-Y. Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019.
31. M. Jas and D. Parikh. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2727–2736, 2015.
32. C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, 2014.
33. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, and D. A. Shamma. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
34. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

35. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
36. 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.
37. H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999. ACM, 2017.
38. M. R. Morris, J. Johnson, C. L. Bennett, and E. Cutrell. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 59. ACM, 2018.
39. K. Papineni, S. Roukos, T. Ward, and W.-J. 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.
40. G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On*, pages 2751–2758. IEEE, 2012.
41. G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016.
42. H. Petrie, C. Harrison, and S. Dev. Describing images on the web: A survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)*, 71, 2005.
43. C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.
44. 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.
45. E. Salisbury, E. Kamar, and M. R. Morris. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of HCOMP 2017*, 2017.
46. E. Salisbury, E. Kamar, and M. R. Morris. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing. In *IJCAI*, pages 5349–5353, 2018.
47. K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. *arXiv preprint arXiv:1810.10665*, 2018.
48. G. Srivastava and R. Srivastava. A survey on automatic image captioning. In *International Conference on Mathematics and Computing*, pages 74–83. Springer, 2018.
49. A. Stangl, M. R. Morris, and D. Gurari. " person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
50. R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

51. L. Von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 79–82. ACM, 2006.
52. 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.
53. 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.
54. 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.
55. X. Yang, K. Tang, H. Zhang, and J. Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.
56. Y. Yoshikawa, Y. Shigeto, and A. Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017.
57. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
58. 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.
59. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.
60. C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688, 2013.