Dimensional alt text: Enhancing Spatial Understanding through Dimensional Layering of Image Descriptions for Screen Reader Users

Jaemin Cho
School of Visual Arts, New York, USA
jaemindesign@gmail.com

Hee Jae Kim
Independent Researcher, Washington, USA
happyhj@gmail.com

ABSTRACT

Over the past decade, there has been a significant improvement in the quality of images we see on the web, and image processing technologies such as monocular depth estimation are opening up new possibilities for various applications. However, despite these developments, how we utilize image descriptions for image accessibility has remained stagnant since alt text was introduced with HTML 2.0 in 1995. This paper presents the concept of Dimensional alt text, which enables users to navigate image descriptions through three-dimensional layers: the foreground, middle ground, and background. Our research findings suggest that providing space for image descriptions on each dimensional layer can assist users in building a mental image of the photo, resulting in better spatial understanding. Our discussion for future work is to extend the use case of the prototype to a broader range of users and investigate a hybrid authoring model that combines human authorship with AI assistance.

CCS CONCEPTS

• **Human-centered computing** → Accessibility; Accessibility technologies.

KEYWORDS

Accessibility, Visual Impairment, Screen Readers, Image Description, Alt Text, Dimensional Alt Text, Depth Map, Alt Text Authoring, Inclusive Design

ACM Reference Format:
Jaemin Cho and Hee Jae Kim. 2023. Dimensional alt text: Enhancing Spatial Understanding through Dimensional Layering of Image Descriptions for Screen Reader Users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544549.3585706

1 INTRODUCTION

Digital images brought about a revolution in the way we capture, share, and consume visual information. Alt text, or alternative text, is a short description of an image that is added to the HTML code of a website, and it is read aloud by screen readers for visually impaired users. Despite the drastic progress in digital image quality, Alt text still uses a single line of text to describe images since its inception in 1995. This poster aims to explore new solutions that will enhance the accessibility of digital images for visually impaired individuals by improving spatial understanding of images with multiple subjects in multiple layers.

In 2017, more than 85% of all photos taken annually were captured on smartphone camera [1], and by 2023, it is estimated that the share will rise by over 93% [2]. The technological advances in smartphone cameras have significantly impacted the digital images we see every day. One of the most notable changes is the increased resolution in images. Furthermore, the introduction of high dynamic range (HDR) technology has greatly improved digital image quality. HDR enables capturing images with great detail in the highlights and shadows in various complex light settings.

The introduction of relative depth map in image analysis has the potential to change the way we understand and interact with digital images. A depth map is a 2D image that encodes the distance of each pixel in the image from the camera, and it effectively separates different layers and elements in the image. By using depth map information, we can quickly identify and isolate different parts of an image, such as the subject of a photograph or the background. The technology is particularly useful for tasks such as image editing, where separating the subject from the background can save significant time and effort.

We see possibilities for utilizing relative depth map to enhance the accessibility of details in digital images for visually impaired individuals. By separating the foreground, the middle ground and the background and providing descriptions on each layer, we can help screen reader users to navigate images three-dimensionally. This interactive interface can help screen reader users build a three-dimensional mental image and be aware of all the details in the images.

2 RELATED WORK

This section reviews a set of references suggesting that alt text and its practice should offer users more control given their diverse preferences. We also investigated research indicating that screen reader users want spatial cues in image descriptions. Additionally, to guide our future work, we conducted a review of the relevant literature, which underscores the critical role of a human author in providing essential context for screen reader users.

2.1 More control for users

There is debate over whether to provide descriptions for decorative images on the internet. While it is a common practice among websites to suggest that decorative images be assigned null alt text,

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585706

Léoni Watson argues that even decorative images should have alt text. According to Watson, screen reader users should have the option to skip around the content, just like sighted users [3].

Morris et al. found that the importance of the progressive detail alt text which gives the user the control over information. They viewed the standard alt text to be passive since there is no interface that gives user control to interact with. To address this limitation, they proposed the progressive alt text which is designed to give the user control over the level of detail received based on their interest in the image [4].

2.2 Needs for spatial cues in image descriptions

Jung et al. discovered from their research what participants want to construct mental images of photos when they view images. In the study, several participants spoke of the need for understanding how the elements of a photo were spatially structured so they could "put things together three-dimensionally" or "structurally imagine" the contents. One participant suggested that consistency in the directionality of image descriptions is important, such as describing an image as moving "from top to bottom" or "from left to right.", and in that sense, the user can more easily construct a mental image of the photo by spatially and logically arranging elements in their heads [5].

2.3 Descriptiveness vs. Conciseness

There are no clear guidelines for the ideal level of details in alt text. Furthermore, personal preferences for the level of detail in alt text may vary. To investigate this matter, Mack et al. [6] conducted an interview on the preferences of screen reader users regarding Descriptiveness versus Conciseness in alt text. The study found out that participants had varying preferences for the level of detail in alt text, ranging from very concise to very detailed descriptions. These results suggest that a one-size-fits-all approach to alt text may not be suitable for all images.

2.4 Human-in-the-loop design

In his article 'AI Is Terrible at Writing Alt Text', Thomas Smith [7] pointed out that even though many companies have turned to Artificial Intelligence to automatically add alt text to images, the resulting text sometimes can end up being inaccurate, offensive, or overly generic. Furthermore, AI-generated alt text typically fails to convey the emotional content of the image.

Mack et al. [6] found out that knowing the context is the key to generating high-quality alt text. The study further proposed that authors can be encouraged to include information that only they know over other types of information, then purely visual descriptions can be created by crowd workers or AI system.

Based on prior research, this research poster focuses on the interactive aspect of alt text interfaces which provides spatial cues to help screen reader users more easily build a mental image of the visual content.

3 LAYERING IMAGES

3.1 Image selection

We conducted a pre-user testing survey to collect more insight into the criteria for selecting sample images. Participants showed a preference for detailed descriptions of people over pets, products, fashion, architecture, and nature. Based on the obtained criteria, we selected images with (1) multiple subjects on multiple dimensional layers and (2) human interactions. For user testing, three sample images were selected: a sports photograph, a self-portrait photograph with a landscape background, and a landscape photograph with people.

3.2 Calculating relative depth

In order to create a greyscale depth map from images, we used the MiDaS Gradio Demo on the Pytorch framework. The demo is built on a model that is pre-trained on large-scale synthetic datasets and fine-tuned on a diverse range of real-world datasets. Thus, it can determine the relative depth of each pixel in an image.

A greyscale depth map image generated by MiDaS Gradio Demo is a visual representation of the relative depth of the objects in a scene. In this representation, the furthest objects appear as solid black, and the nearest objects appear as solid white. The intensity of the grayscale value in between represents the distance of objects in the scene (Figure 1).

3.3 Separating layers

When creating dimensional layers based on relative depth, the process starts by determining the range for layering. We reviewed hundreds of images in greyscale to identify the appropriate range of colors for layering in various compositions.

The first layer, Layer 1, is defined by an RGB value range of (255,255,255) to (180,180,180). This range corresponds to the lightest areas in the grayscale image and represents the closest elements in the scene to the viewer. The second layer, Layer 2, is defined by an RGB value range of (180,180,180) to (100,100,100). This range corresponds to the mid-tone areas in the grayscale image and is farther away than those in Layer 1 but still represents elements closer to the viewer than those in Layer 3. The third layer, Layer 3, is defined by an RGB value range of (100,100,100) to (0,0,0) and represents the most distant elements in the scene. This range corresponds to the darkest areas in the grayscale image.

It is important to note that not all photos may have the same number of layers. Some images may be missing either the second layer or both second and third layers, depending on the composition and visual elements of the photo.

3.4 Labeling of interfaces

For intuitive comprehension of the concept, it is crucial to use familiar terminology to label user interfaces. To align with the composition elements of photography, the layers were named from front to back, as "Foreground", "Middle Ground," and "Background". Using these terms, it is easy to understand the relationship between different layers and how they create a sense of depth in an image. Thumbnails displayed below the images were named "Dimensional

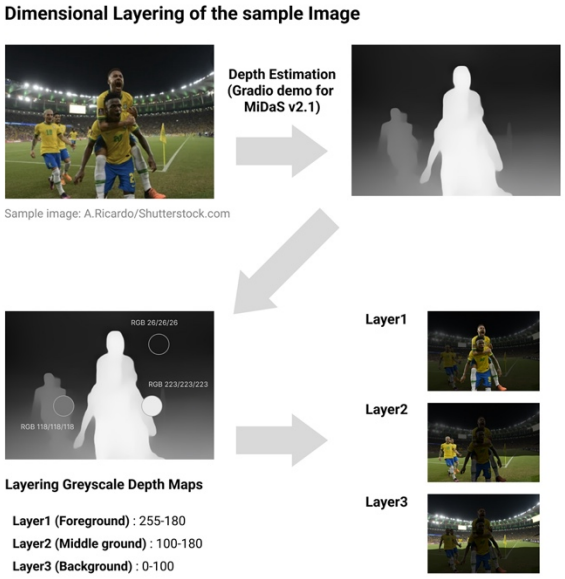


Figure 1: Layering dimensional layers using monocular depth estimation (Sample image: A. Ricardo. 2022. Shutterstock. Retrieved January 10, 2023)

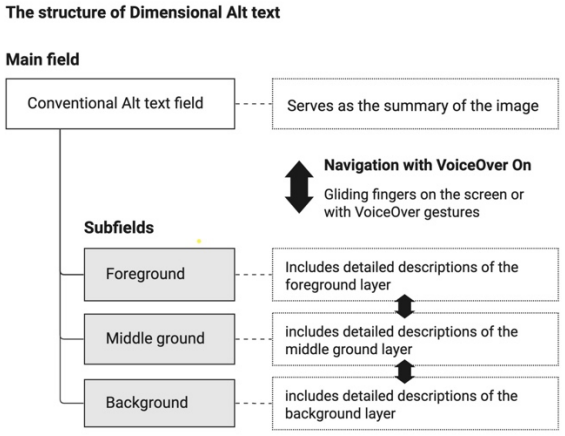


Figure 2: The Dimensional alt text structure consists of conventional alt text as a main field and three subfields for additional descriptive details.

thumbnails". The Dimensional thumbnails (small images representing different layers of an image) allow for a quick preview of text descriptions on different layers with screen readers.

The results from user testing indicate that using clear and familiar terminology helped the user understand the concept quickly and gave them more confidence to interact with the prototype without worrying about failing.

4 PROTOTYPE DESIGN

4.1 Structure

The Dimensional alt text consists of one main field and three dimensional subfields (Figure 2). The main field, the conventional alt text field, serves as a summary of the image. The three dimensional subfields provide space for more detailed descriptions of each layer within the image, allowing users to interact with image descriptions.

4.2 Thumbnails

The dimension thumbnails are located below the sample image (Figure 3). These thumbnails look like tabs or buttons but are images with accompanying alt texts. To help screen reader users quickly identify the content of each thumbnail, alt texts for these images start with "Foreground", "Middle ground", or "Background". Testers can easily browse back and forth through thumbnails and stay only on the dimensional layer that they are interested in.

4.3 Comparison

The prototype aims to compare usability between the conventional alt text and Dimensional alt text for three images. The prototype features a global navigation with six tabs. The screens labeled 1A,2A, and 3A show sample images with the conventional alt text, and the screen labeled 1B,2B, and 3B show sample images with the Dimensional alt text.

4.4 Two ways to navigate

To ensure accessibility for screen reader users, the prototype was designed and tested for use on Safari with VoiceOver on iOS devices. Screen reader users have two options for navigating through the main image and Dimensional thumbnails.

The first option is to run their finger over the screen while VoiceOver is on. After giving an overview of the main image, a guide instructs users to "Tap the Dimension thumbnails below for more details." The Dimensional thumbnails are located directly below the main image, and this allows users to easily glide their fingers over them to hear descriptions of the foreground, middle ground, and background.

The second option involves using VoiceOver gestures [8]. With VoiceOver activated, users can swipe right to move the focus to the next element on the screen or swipe left to return to the previous element. For example, if the user is currently focused on the main image and wants to move to the Dimensional thumbnails, they can swipe right to move the focus to the first dimensional thumbnail.

5 USER TESTING

5.1 Test environment

We designed a web-based prototype, and participants for the user testing accessed the website using Safari on iPhone, with VoiceOver.

5.2 Participants

Participants were recruited from the assistive technology community on Facebook. A total of five individuals with varying vision conditions were selected to participate in the user test session. This

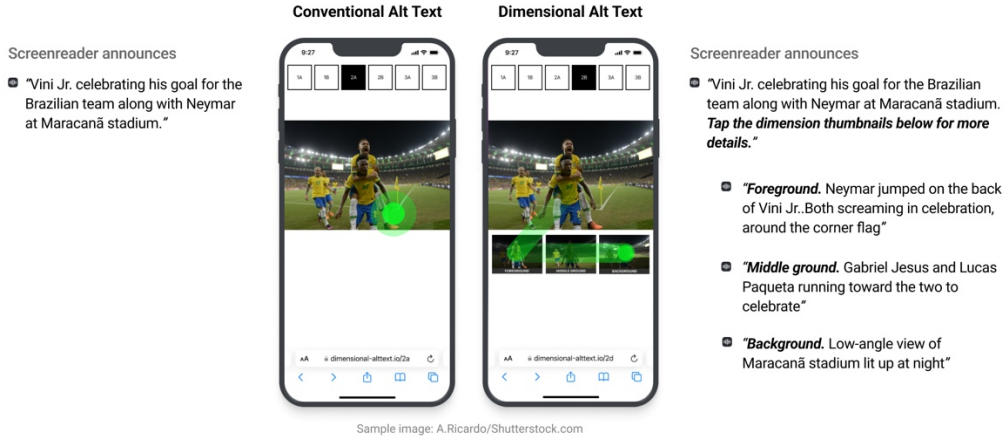


Figure 3: Prototype of Dimensional alt text for comparison with conventional alt text.

sampling strategy aimed to ensure a diverse range of perspectives and experiences to be represented in the evaluation of the prototype.

5.3 Procedure

The user testing was conducted remotely. Participants were given guidance on setting up their test environment and instructed to proceed accordingly. Participants connected their computers and smartphones to Zoom and broadcasted their phone screens. This enabled a more thorough observation of participants' reactions and interactions with the prototype on-screen during the testing process.

6 INSIGHTS

Our user testing revealed key insights of how users found the Dimensional alt text helpful to better understand images three-dimensionally. We characterize three main insights from the feedback of participants: (1) Navigation style, (2) More focus on backgrounds, (3) Cue for the composition.

6.1 Navigation style

P2 noted that using dimensional layers was helpful in creating a three-dimensional mental image of the photograph, commenting "It felt like walking into the photo layer by layer." However, P2 suggested an alternative navigation style that allows users to navigate the image by gliding their fingers on the photo rather than navigating with tabs or buttons below the image.

Other participants remarked that the Dimensional alt text's navigation style aligns with how photos are composed, stating that the primary subject is typically in the foreground of photos. P3 highlighted that in most photos, the navigational order of the Dimensional alt text reflects the importance of the information. This means that users can save time when using screen readers, as they can start reviewing the descriptions from the foreground, where the most important information is usually located.

All participants answered that they could easily comprehend the labeling of subfields without a learning curve, since they are commonly used photography terminology. Participants also reported that they could confidently navigate the user interface even if it was their first time using it.

6.2 More focus on backgrounds

Participants pointed out that Dimensional alt texts can focus on the backgrounds that are mostly ignored in the current alt text descriptions. P1 commented, "A lot of the time, even with magnifying pictures, I don't really understand backgrounds and things going around somewhere else in the picture, even if I magnify to the maximum zoom level."

P5 explained, "For pictures dealing with sports, the image description tends to focus on just what's in the front, or even on the single player. But something else might be happening in the background, and even if the background detail is very basic, that's all someone might need it. The description of the middle ground and the background (in Dimensional alt text) can fill in the rest of what is going on in the picture."

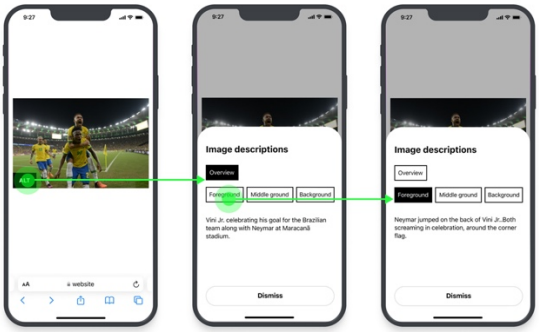
6.3 Intuitive cues for composition

Participants expressed that the number of Dimensional thumbnails shown below the image can make users understand the structure and the angle of the photo at a glance, especially for images that do not have a middle or a background. P4 commented on the screen of tab 3B (Figure 4), "For 3B, I instantly recognize that there is no middle ground. That means that the background is taking up the shot. Maybe it is an upwards angle so that there is no middle ground. The fact that there was no middle ground here did give me that extra information of how the shot was laid out."

P5 suggested even when the background is blurred for the bokeh effect or deleted in the post-processing, if that is stated in the background subfield, it will give a better understanding of how images are structured.

Sample image: Personal Belongings/Shutterstock.com

Figure 4: The image without a middle-ground Dimensional thumbnail due to its composition. (Sample image: Personal Belongings. n.d. Shutterstock. Retrieved January 10, 2023)



Sample image: A.Ricardo/Shutterstock.com

Figure 5: The new proposal based on Twitter's Alt badge and modal sheet approach to include magnifier users in the use case of Dimensional alt text.

7 DISCUSSION FOR FUTURE WORK

In this section, we discuss potential avenues for future research based on the key findings obtained from our study.

7.1 Including magnifier users

Participants who use both screen readers and magnifiers pointed out that the Dimensional alt text can benefit not only the screen reader users but also the magnifier users if the description is provided as text somewhere in the interface. Some participants further proposed expanding the use case for the magnifier users like Twitter's alt text badge case [9].

P1 commented, "In the prototype, only people that use screen readers would be able to access alt texts and benefit from it, whereas there are people like me who don't use a screen reader all the time but still struggle with pictures."

We propose a new design solution, informed by the feedback (Figure 5). This design caters to the needs of magnifier users by providing access to detailed descriptions of each dimensional layer.

7.2 Web standards for depth layer descriptions

We can consider including descriptions of additional layers using web standards, such as the aria-details attribute of the WAI-ARIA property. This attribute can take a set of ID references as its value, which can point to another element containing the description for each layer of the image element. However, because screen reader software has limited support for this feature, users may not be able to benefit from it out of the box. To enhance the user experience, we can modify the existing HTML on the fly to make the depth layer descriptions accessible and navigable while maintaining a connection to the original image. For example, we can reconstruct the HTML content with a web browser extension to add an extra level of navigation.

7.3 Hybrid authoring model

The results of a pre-survey indicate that all five participants are dissatisfied with the quality of AI-generated image descriptions. P2 stated that AI "does more harm than good" since it cannot comprehend why the author chose a specific photo. Mack et al. [6] suggested that context is crucial for creating accurate descriptions and further argued that authors could be encouraged to provide critical information about why an image was chosen.

Based on the participants' feedback and literature review, we propose the hybrid authoring model that combines the use of generative AI with the Human-in-the-Loop approach for Dimensional alt text as a future direction. This approach aims to reduce the author's workload by using AI to generate descriptive alt text for each dimensional layer while still including crucial contextual information in the alt text.

In the hybrid authoring model, the author initiates the process by providing essential contextual information or specific details that are not expected to be identified by the Machine-based system. Subsequently, AI combines the human-supplied context with computationally analyzed properties (including depth map) while generating detailed descriptions for each subfield. The author can review the AI-generated descriptions after all subfields have been completed (see Figure 6 for a visual representation of this process).

8 CONCLUSION

Our research highlights the need for improvements to the current practice of alt text for digital images, particularly given the significant advances in image quality in recent years. We proposed that image descriptions can benefit from utilizing advancements in image processing technology, such as relative depth maps. To explore new possibilities in alt text, we developed a prototype called "Dimensional alt text". By providing descriptive details and spatial cues for each dimensional layer, this approach enhances the user's spatial understanding of the image, especially images with multiple subjects on multiple layers. We gathered insights from five screen reader users on the prototype and proposed potential future work. As future directions for research, we recommend expanding the implementation of Dimensional alt text to a broader audience such as screen magnifier users. Additionally, we propose a hybrid authoring model that combines generative AI with the human-in-the-loop approach to create contextual alt text, minimizing the burden on

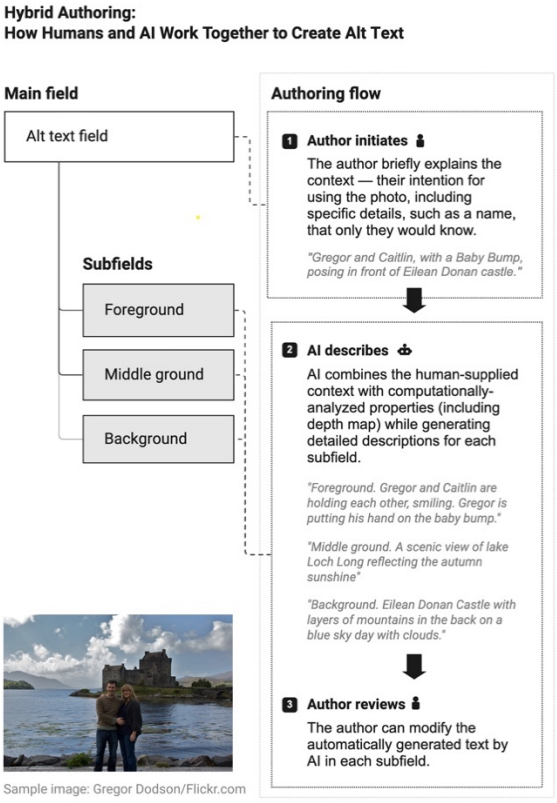


Figure 6: Workflow of Hybrid Authoring Model. (Sample image: Gregor Dodson. 2008. Flickr. Retrieved January 10, 2023. CC BY-SA 2.0)

authors. Overall, our research aims to leverage current technologies to enhance the accessibility of digital images, with a specific focus on enabling screen reader users to access descriptive details in images with multiple subjects on multiple layers, which was previously challenging for them.

REFERENCES

[1] Felix Richter (August 31, 2017). Smartphones Cause Photography Boom. Retrieved January 15, 2023, from https://www.statista.com/chart/10913/number-of-photos-taken-worldwide/

[2] Ed Lee (June 10, 2021). 2021 Worldwide Image Capture Forecast: 2020 – 2025. Retrieved Jan 15, 2023, from https://riseaboveresearch.com/rar-reports/2021-worldwide-image-capture-forecast-2020-2025/

[3] Léonie Watson (June 2011). Text descriptions and emotion rich images. Retrieved January 15, 2023, from https://tink.uk/text-descriptions-emotion-rich-images/

[4] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 59, 1–11. https://doi.org/10.1145/3173574.3173633

[5] Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S. Ackerman. 2022. "So What? What's That to Do With Me?" Expectations of People With Visual Impairments for Image Descriptions in Their Personal Photo Activities. In Designing Interactive Systems Conference (DIS '22). Association for Computing Machinery, New York, NY, USA, 1893–1906. https://doi.org/10.1145/3532106.3533522

[6] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21). Association for Computing Machinery, New York, NY, USA, Article 23, 1–14. https://doi.org/10.1145/3441852.3471207

[7] Thomas Smith (May 2021). AI Is Terrible at Writing Alt Text. Medium. Retrieved March 3, 2023, from https://tomsmith585.medium.com/ai-is-terrible-at-writing-alt-text-e79b0c4ecf51

[8] Apple Inc. 2021. iPhone User Guide. Retrieved March 3, 2023, from https://support.apple.com/en-gb/guide/iphone/iph3e2e2281/ios

[9] Lauren Race (2012). Twitter Alt Text Features. laurenrace.com. Retrieved March 3, 2023, from https://laurenrace.com/design/twitter-alt-text-features-design/