Sketchy Collections: Exploring Digital Museum Collections by Drawing via CLIP

Polo Sologub, Rebecca Fiebrink
Creative Computing Institute
University of the Arts London
London, UK
{p.sologub, r.fiebrink}@arts.ac.uk

Abstract

In recent years, digital museum collections have made it possible for everyone to discover cultural heritage (CH) online. However, that does not mean that they are engaging or fun for casual users to explore. In this paper, we develop a web interface that lets users search and compare three museum collections by drawing images. We describe our approach of using CLIP as a feature extraction model for a Sketch-Based Image Retrieval (SBIR) model based on museum tags. Through qualitative experiments and a user study, we demonstrate that the model performs well in a CH context with interesting results and that the interface enables playful search and serendipitous discoveries.

1 Introduction

In recent years, museums have been digitising their collections so that they can be accessed online. However, a lot of their collections websites are difficult or not interesting to explore for casual users due to traditional catalogue-style interfaces with key word search which requires users to know the collection or what to look for [1]. This has raised an interest in innovative interfaces and computer vision for exploring and discovering collections in new ways [2]. For example, "generous interfaces" with rich, browsable overviews of large collections were proposed by [3] as an alternative to the search box. But is the search box really that restrictive? We believe that alternative forms of the search box, such as searching by drawing, have the potential to facilitate playful and serendipitous exploration for casual users.

Sketch-Based Image Retrieval (SBIR) is concerned with using query sketches to retrieve similar images. It has found limited use outside computer vision research as part of interfaces and even less so in a cultural heritage context (with the exception of [4] and [5]). SBIR has specific challenges, such as the abstract nature of human drawings, as drawing skills and interpretations of the sketches can vary [6]. Recently, the challenges of creating and annotating large-scale sketch datasets have motivated the development of zero-shot SBIR (ZS-SBIR) [6]. This method introduces zero-shot learning capabilities to SBIR, allowing the retrieval of images from unseen categories in hopes of increasing the generality and usability of real-world applications [7, 6, 8]. This line of ZS-SBIR research mostly involves training models from scratch, although leveraging CLIP has recently gained popularity [9]. Here, we chose CLIP [10] for feature extraction to facilitate quicker prototyping for an interface, and for its semantic nature, as it can be combined with the existing vocabulary available as metadata in museum databases.

Preprint. Under review.

2 Approach

2.1 System

Building on [11], which uses CLIP for text-to-image retrieval, we created an image-to-image retrieval system. Given a list of textual labels encoded with the prompt “an image of [label]”, we used CLIP as a feature extraction model to turn both the query and dataset images into text embeddings of the most likely labels. Finally, we used cosine similarity to find the most similar images based on the full text embeddings (Fig. 1). While the system can take any type of image as a query and the interface includes an image upload option, this paper focuses on using sketches as queries.

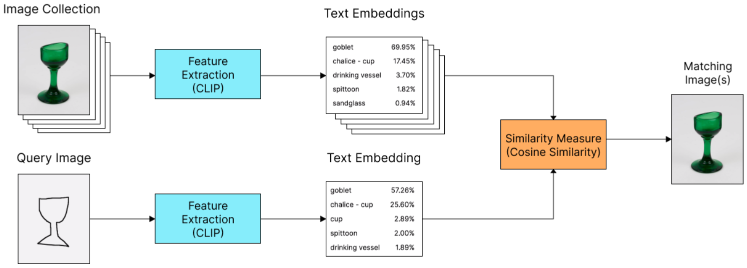


Figure 1: Explanation of the image retrieval system.

We used the system on three digital museum collections: The Metropolitan Museum of Art, Science Museum Group, and Cooper Hewitt, Smithsonian Design Museum (Table 1). The museum collections were chosen for their contrasting subjects, availability of public APIs including images and availability of lists of tags or similar vocabularies for the textual labels. Tags were preferred because an early experiment using a mixed list of labels of the types “tag”, “medium”, “country”, and “dimension” [12] revealed that tags were almost exclusively among the top five predicted labels for an image. Another early experiment showed that a dataset of 2,500 images was too small and led to a lack of variation in the results. Therefore, we aimed to collect at least 10,000 images per dataset.

Table 1: Datasets collected from each museum.

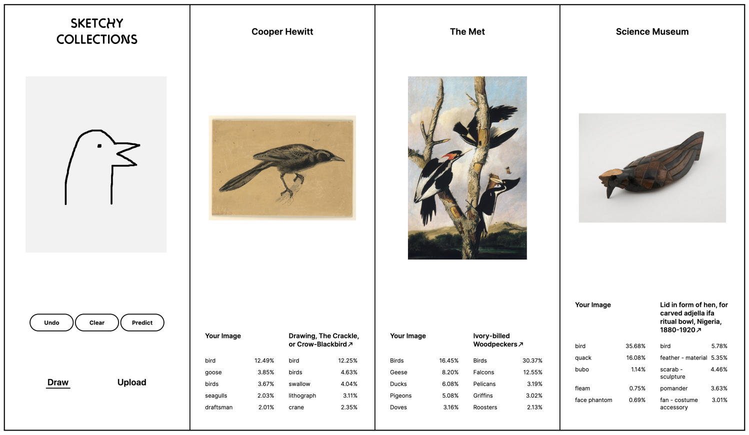
Museum	No. of objects (online)	No. of objects (collected)	Type of labels	No. of labels (collected)
Cooper Hewitt, Smithsonian Design Museum	193,698	41,833	Tags	3,681
The Metropolitan Museum of Art	>492,000	32,885	Tags	1,162
Science Museum Group	>380,000	40,979	Taxonomy	1,531

We used the museum APIs [13, 14, 15] to create two datasets for each collection: a list of tags (or similar) and a list of CC0-licensed [16] museum objects including the URL for their primary thumbnail and their object ID. The image URLs were used to download the images for pre-processing and the resulting feature embeddings were stored in NumPy files. The object IDs were used to later retrieve these images from the APIs and any metadata to be displayed alongside them in the interface.

2.2 Interface

We built a web interface using vanilla JavaScript and Flask. The back-end contains a Python script with the IR model and the data required to run it: NumPy files of the feature encodings and object IDs, and CSV files of the tags. The layout is divided into four columns: the first column is the input section which contains the drawing tool. In addition to the drawing option, the interface includes an image upload option that uses the same IR system. The drawing tool was made using Canvas API. It lets the user make simple black line drawings. Additionally, an “Undo” button lets the previous stroke be erased, and a “Clear” button removes the entire drawing. The remaining three columns are the output columns, one for each museum. Both the drawing and image upload tools have a “Predict” button. This sends the drawing or image as a query to the model and returns the top match for each museum. When the user clicks on the “Predict” button, the image of the top match appears in each of the output columns. Underneath each output image, two sets of top 5 tags are displayed. On the left, titled “Your Image”, are the top 5 tags that were generated for the query image. On the right, the title of the artwork is displayed and the top 5 tags that were generated for it. The title also contains an external link to the catalogue entry of the artwork in the digital catalogue of the museums, giving the user the option to explore the artwork further (Fig. 2).

Displaying the different tags and images of the museums side by side allows the user to compare the predicted tags with the images, as well as compare the different types of tags between the museums. This invites the user to explore different aspects: a glimpse into the museum collection via the artworks, into the workings of the AI when comparing the tags and images, and into the structure of the museum by showing the different types of tags.



Your Image		Drawing, The Crackle, or Crow-Blackbird↗	
bird	12.49%	bird	12.25%
goose	3.85%	birds	4.63%
birds	3.67%	swallow	4.04%
seagulls	2.03%	lithograph	3.11%
draftsman	2.01%	crane	2.35%

Your Image		Ivory-billed Woodpeckers↗	
Birds	16.45%	Birds	30.37%
Geese	8.20%	Falcons	12.55%
Ducks	6.08%	Pelicans	3.19%
Pigeons	5.08%	Griffins	3.02%
Doves	3.16%	Roosters	2.13%

Your Image		Lid in form of hen, for carved adjella ifa ritual bowl, Nigeria, 1880-1920↗	
bird	35.68%	bird	5.78%
quack	16.08%	feather - material	5.35%
bubo	1.14%	scarab - sculpture	4.46%
fleam	0.75%	pomander	3.63%
face phantom	0.69%	fan - costume accessory	3.01%

Figure 2: Screenshot of the interface showing the drawing tool.

3 Evaluation

3.1 System

To evaluate the system on the three museum collections, we used sketches from the TU-Berlin dataset, which contains 20,000 sketches by laypeople across 250 categories of everyday objects [17]. We chose one random sketch from each category to use as queries and conducted a thematic analysis of the results. We looked at the top 3 matches for each museum including the top 5 tags for each query and match. Because the goal of the project was to create an engaging interface, we did not look for quantitative accuracy. Instead, we aimed to determine if the results made some sense (as opposed to being random) and if there were any interesting and surprising results. We found that it was helpful to compare the predicted tags for the three collections to get a sense of what is predicted based on the limitations of the vocabulary and the museum collections. This would often reveal a theme even if there was a seemingly inaccurate image.

The system returned consistent results of matching objects across the three museum collections for sketches of common objects and animals, e.g. vases (Fig. 3), mugs, and dogs. In cases where the sketches looked more ambiguous, such as a sketch of a doughnut (Fig. 4), the results showed a variety of different objects that looked similar based on the shape of the sketch, e.g. a painting of a plate.

The model provided fascinating results due to its semantic embedding space. Results such as for the sketch of a snowman wearing a hat (Fig. 5) showed the range of details the model could pick up on in a single sketch. Here, it returned a Christmas card, and two objects that it recognised as ice and a hat. Due to the model's semantic nature, sketch queries of modern objects such as a handheld console led to surprising connections to historical objects, e.g. a set of buttons (for clothing) and a set of dice (Fig. 6). Another compelling theme was where the results provided objects in response to the query rather than just direct matches: for instance, sunglasses against the sun, and a piece of bamboo for a bear (Fig. 7).

In some cases, we used several sketches from the same category to test the consistency of the model across different queries. We observed that the model provided similar matches for sketches that looked similar, whereas different looking sketches from the same category often yielded different results.

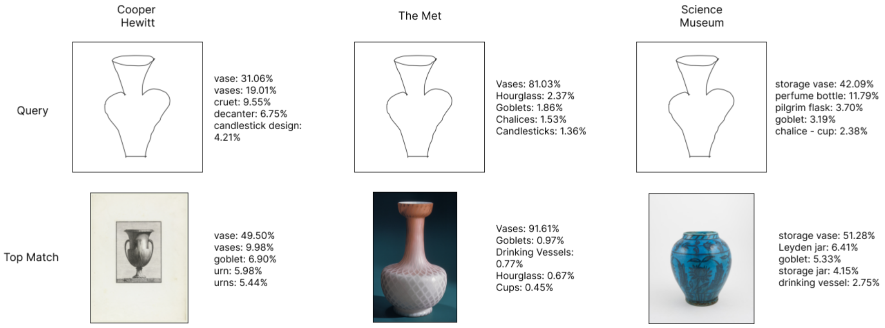


Figure 3: Results for a sketch of a vase showing the top match from each collection. For each collection, the query sketch and matching image are displayed alongside the top 5 predicted textual labels.

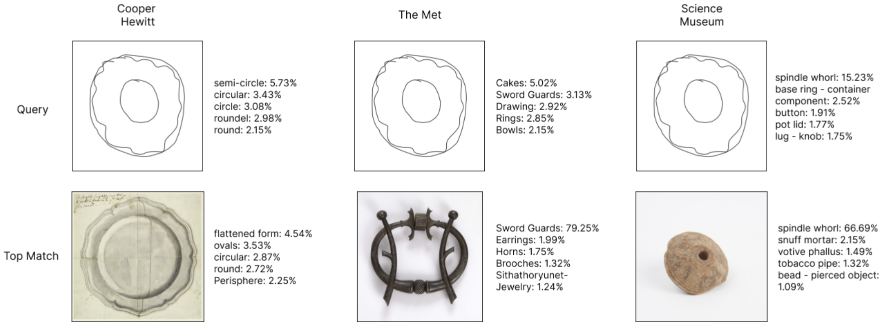


Figure 4: Results for a sketch of a doughnut showing the top match from each collection.

3.2 Interface

To evaluate the interface, we conducted a user study with four laypeople who visit museums for leisure. The aim of the study was to discover possible search behaviours to explore whether and how the search tool can allow for playful, serendipitous interactions. Each of the user study sessions was conducted in person and took 1 hour. It consisted of an introduction to the interface followed by a free exploration of the interface by the participant in private followed by a semi-structured interview.

Overall, the participants described the tool as fun and interesting. P3 expressed that, although not knowing what to look for initially, she ended up feeling engaged in an open-ended and playful exploration: *"It was fun! I probably could have just kept doing that for ages, to be honest. It was quite engaging in a way. I was a bit like, 'I don't know what I'm going to do with this', but then once I was actually playing, I found it fun!"* P2 shared the sentiment of open-endedness, telling me that she did not want to stop experimenting with the tool. Similarly to P3, P1 tends to struggle with knowing what to look for. He found that the unexpected and sometimes seemingly random nature of the drawings' results was helpful in overcoming this hurdle: *"Basically, the look of the drawing and not knowing what you were going to get. That would be why I'd use something like that, as I often do have a hard time making up my mind. So, for something like that to choose for me at random, that's fun."* Not only did he enjoy the possibility of random outputs, but also the possibility of putting in something random: *"I guess I like this tool because it is a little bit of a wild card: put in something random and then it's like the 'I'm feeling lucky' search function on Google."*

Interestingly, there seemed to be a personal and active interaction with the drawing tool. When drawing, P4 wanted to collaborate with the model and be understood by it: *"With the drawing I found that I wanted it to be like a collaboration, like I really wanted it [the program] to understand. Like, 'this is a cell'. But I wasn't mad that it didn't."* P1 wanted to give the model something more up to its interpretation. Other participants wanted to trick the model by drawing images that they thought were ambiguous enough for the model to be confused. Feelings about their drawing skills also played a role. In some cases, they were not confident that a drawing of theirs would be understood but were surprised that it did: *"It surprised me how good it is at getting what the thing is that you were doing,*

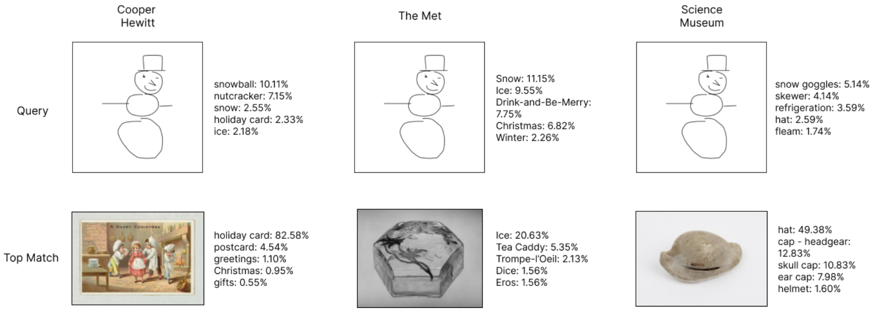


Figure 5: Results for a sketch of a snowman showing the top match from each collection.

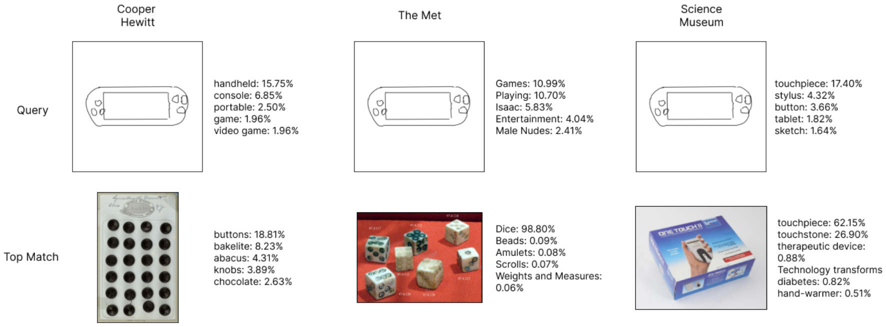


Figure 6: Results for a sketch of a handheld console showing the top match from each collection.

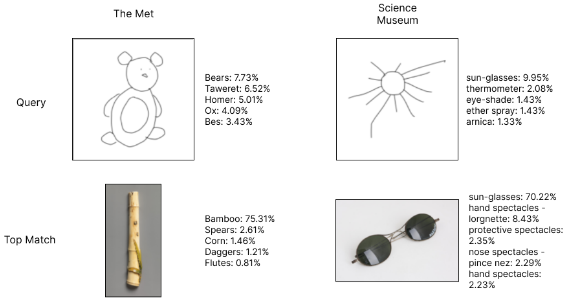


Figure 7: A result for a sketch of a bear and a result for a sketch of the sun.

even when it looks like shit." [P3]. In one case, it was the opposite; P1 was offended that he did not see accurate results for his drawing.

When providing feedback on the interface, some participants stressed that they enjoyed the novelty of the drawing tool and that it fits the art theme. Some participants liked comparing the three museums: P2 found it *"cool to see the differences of the three museums"* whereas P4 liked *"the order of the museums"* displayed in the interface for aesthetic reasons. The participants were curious about the tags and actively took them into account when looking at the results. In fact, some participants were interested in seeing more than the top 5 tags: *"I'd quite like to see the other percentages. See the top five, but [I would like] to see the whole list. I don't know what I'd do with that information, just out of interest."* [P1].

Some participants wished for more options to explore the wider context of the collections. P2 suggested that in order for her to be more curious about the collections rather than just playing with the tool, she needed to see more about them, for example, via a button that lets the user explore the wider collection with the match as a starting point: *"Maybe to have something else at the bottom when you get these results. Then you can go on like 'explore this collection' or something like that. Or see things that are in this collection but maybe not so related to the other image."* Or have a more generous results page with more matches and descriptions: *"Like make it very visual. Like the main* *thing [result] and then have other ones. Like three or four."* P3 also wanted to see more matches without having to submit a new query because the results often did not match his interests: *"If it was almost like a slot machine, if you could like click hold or keep spinning it. I think like I went in and I wanted specific results for this one thing but then have to submit it again; rather than that have somewhere I could just keep shuffling them until I landed or pick and choose options that I liked."*

4 Discussion and Conclusion

By qualitatively evaluating the results for different sketches, we came to the conclusion that CLIP is a promising feature extraction model for SBIR in a CH context when being provided domain-specific vocabularies, even without any fine-tuning. Previous research has shown that fine-grained classification of artworks is a difficult task, especially when it comes to the detection of objects within paintings, rather than just labelling them "painting" [18]. However, the model performed very well at recognising what artworks may depict, no matter what medium. The results are consistent with [10], as they show that CLIP generalises well across a large variety of museum objects across the science, design, and art domains from different historical periods. The semantic nature of the CLIP-based SBIR model led to fascinating results. Not only could it return exact matches for sketches (Fig. 3) or ones with similar shapes (Fig. 4), it could reveal objects with related themes (Fig. 5), or provide "tools" in response to them (Fig. 7). Oftentimes, the results showed the contemporary lens of the model when looking at historical data (Fig. 6).

It was interesting to observe how the results illuminated both the scope and limitations of the museum collections, their vocabulary, and the way CLIP has been trained to see the world. When comparing the images and predicted tags for the three collections, we were able to understand better what CLIP saw in the images, even when the matches seemed inaccurate at first. Doing this also gave us an intuitive way of learning about the museum collections. For example, when receiving consistent and "accurate" results for sketches such as a vase (Fig. 3), it showed us that this type of object is well-represented in all three museums. When results were not as obvious, we were invited to look more closely at the tags alongside the images, which revealed the language used behind the scenes at a museum.

Even though the user study was limited to four participants, we found that their responses suggested strongly that the interface enables playful search and serendipitous discoveries. The participants reported that they were compelled to play with the tool for an extended amount of time and that they were surprised by different aspects of the tool, such as the ability of the model to recognise their sketches and return unexpected results. They expressed that the tool helped them overcome the initial hurdle of deciding what to look for in a collection. The participants were intrigued by the AI system and showed interesting behaviours when engaging with it, from trying to collaborate with it and be understood to trying to trick it. However, some users raised that they wanted more from the interface to explore the museum collections further rather than just playing with the tool, for example, seeing more matches and tags and the wider context of a collection. This suggests that search tools on their own may indeed, as suggested by [3], be not enough to explore CH collections fully. Further work may address this limitation by expanding the interface with more matches and options to explore the wider context of the museum collections, e.g. through overviews.

5 Ethical Implications

Due to working with a neural network, we were expecting a large chance of ethically questionable results from the retrieval system, especially as it was stressed by [10] that the results can be hard to predict and depend a lot on the prompts. The risk seemed lower when using sketches as queries, as risky results only appeared when testing the system with images, such as a stock photograph of two women from Unsplash [19] being labelled as "prostitutes". This also shone a light on the fact that museum tags are inherently biased and therefore impact the results. It made clear that there is ingrained bias in all parts of the system – the data and the AI model - which echoes the findings of [20]. Because of results like these, we expected there to be more risk for the participants during the user study and warned them about unexpected and potentially offensive results. While there was only one instance where a result made a user uncomfortable, it remains significant.

References

[1] David Walsh and Mark M Hall. Just looking around: Supporting casual users initial encounters with digital cultural heritage. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks at ECIR 2015*, 2015.

[2] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE Transactions on Visualization and Computer Graphics*, 25(6): 2311–2330, 2019.

[3] Mitchell Whitelaw. Generous interfaces for digital cultural collections. *Digital Humanities Quarterly*, 009(1), 2015.

[4] Google Creative Lab. Quick, draw!, 2017. URL `https://experiments.withgoogle.com/quick-draw/`.

[5] Google Creative Lab, Google Arts & Culture Lab, and IYOIYO. Draw to art, 2018. URL `https://experiments.withgoogle.com/draw-to-art`.

[6] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2174–2183, 2019.

[7] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Computer Vision – ECCV 2018*, volume 11208, pages 316–333. Springer International Publishing, 2018.

[8] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765 – 2775, 2023.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[11] Vladimir Haltakov. Unsplash image search, 2023. URL `https://github.com/haltakov/natural-language-image-search`.

[12] Chenyang Zhang, Jong-Chyi Su, and nikhil trivedi. imet collection 2021 x aic - fgvc8, 2021. URL `https://kaggle.com/competitions/imet-2021-fgvc8`.

[13] Cooper Hewitt, Smithsonian Design Museum. Access our data with the cooper hewitt api. URL `https://apidocs.cooperhewitt.org/api-home/`.

[14] The Metropolitan Museum of Art. The metropolitan museum of art collection api. URL `https://metmuseum.github.io/`.

[15] Science Museum Group. Using our collection api. URL `https://www.sciencemuseumgroup.org.uk/our-work/our-collection/using-our-collection-api`.

[16] Creative Commons. CC0. URL `https://creativecommons.org/public-domain/cc0/`.

[17] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[18] Eva Cetinic and James She. Understanding and creating art with ai: Review and outlook. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2), 2022.

[19] Unsplash. Unsplash. URL `https://unsplash.com/`.

[20] Elena Villaespesa and Oonagh Murphy. This is not an apple! benefits and challenges of applying computer vision to museum collections. *Museum Management and Curatorship*, 36(4):362–383, 2021.