Visual Resources

an international journal on images and their uses

ISSN: 0197-3762 (Print) 1477-2809 (Online) Journal homepage: www.tandfonline.com/journals/gvir20

There Is a Digital Art History

Leonardo Impett & Fabian Offert

To cite this article: Leonardo Impett & Fabian Offert (2022) There Is a Digital Art History, Visual Resources, 38:2, 186-209, DOI: 10.1080/01973762.2024.2362466

To link to this article: https://doi.org/10.1080/01973762.2024.2362466

Published online: 07 Aug 2024.

There Is a Digital Art History

Leonardo Impett and Fabian Offert*

In this paper, we revisit Johanna Drucker's question, "Is there a digital art history?" – posed a decade ago in this journal – in the light of the emergence of large-scale, transformer-based vision models. While more traditional forms of neural network have long been part of digital art history, and digital humanities projects have recently begun to use transformer models, their epistemic implications and methodological affordances have not yet been systematically understood. We focus our analysis on two main aspects that, together, seem to suggest a coming paradigm shift towards a "digital" art history in Drucker's sense. The visual-cultural repertoire newly encoded in large-scale vision models has an outsized effect on digital art history. The inclusion of significant numbers of non-photographic images allows for the extraction and automation of a much wider gamut of visual logics. Large-scale vision models have "seen" huge swathes of online visual culture, biased towards an exclusionary visual canon; they continuously solidify and concretize this canon through their already widespread application in all aspects of digital life. We use a large-scale vision model to propose a new critical methodology that acknowledges the epistemic entanglement of neural network and dataset. We propose that, in reading a corpus of visual culture through a neural network, we are always also doing the reverse. Digital art history is here, but not in the way we expected: rather, it has emerged as a crucial route to understanding, exposing, and critiquing the visual ideology of contemporary AI models.

Keywords: Digital art history; Artificial intelligence; Computer vision

Introduction

In 2013, in this journal, Johanna Drucker asked: "Is There a 'Digital' Art History?" Her provocatively titled article suggests an important difference between "digitized" and "digital" art history. The former refers to the "making digital" of visual culture, a somewhat successful set of practices including digitization, database creation, or virtual exhibitions. But there has been "no research breakthrough" so far, Drucker argues, and we have yet to witness "a convincing demonstration that digital methods change the way we

*Both authors contributed equally to this article.

Visual Resources, Volume 38, Number 2, June 2022
ISSN 0197-3762

https://doi.org/10.1080/01973762.2024.2362466

understand the objects of our inquiry."[1] A "clear distinction has to be made between the use of online repositories and images, which is digitized art history, and the use of analytic techniques enabled by computational technology that is the proper domain of digital art history."[2] Technical shifts have not led to epistemic ones that might "reconfigure our fundamental understanding of what constitutes a work of art."[3] While we can always have more convenience through digitization, a digital art history that enables new kinds of questions has yet to emerge.

Drucker's argument, of course, mirrors the fundamental methodological question of the digital humanities in general – what is the "surplus value"[4] of the digital? One answer, epitomized in the notion of "distant reading"[5]/"distant viewing"[6] is scale. But the rare works of digital art history that have focused on intrinsically large-scale problems – Diana Greenwald on the economic history of nineteenth-century Western art production,[7] or Matthew Lincoln on networks of early modern Dutch and Flemish printmaking[8] – have tended to do so outside the realm of the (digital) image. In the past decade, digital art history – or perhaps more accurately, digital visual studies – has often been arrested by a very basic question, which turned out to be a significant technical challenge: what is *in* an image?

Maybe the divergence between digital art history and other forms of digital humanities lies in what we could call the *Laocoön* problem of computation: the affordances of (electronic) images are different from those of (electronic) texts. Computationally, the two are almost diametrically opposed. There is no equivalent to the discrete "tokens" (letters or glyphs) of text-based digital approaches in the visual domain, and there is no widely accepted hierarchy of elements (the paragraph, the sentence, the word). Sentences end with a period. But where do image-objects end, exactly? Most texts have a clearly defined vocabulary – some number of words or subwords, with a reasonable upper bound. In images, potential "vocabularies" are ambiguous, infinite, and not tied to any common superset. And even if images share a common – for instance iconographic – visual vocabulary, no instance of a "word" is like the other.[9] The ingenuity of Lev Manovich's pioneering work on image sets qua *style spaces*[10] is to sidestep this question completely: digital images are pixels, and pixels have measurable properties like color, brightness, and entropy. At the same time, this approach limits our inquiry to the kind of phenomena measurable at the pixel level. The challenges for digital methods in art history, then, seem different – and perhaps more fundamental[11] – from those in other branches of the digital humanities. Johanna Drucker's question still stands, ten years after it was first proposed: has there been a "digital" art history?

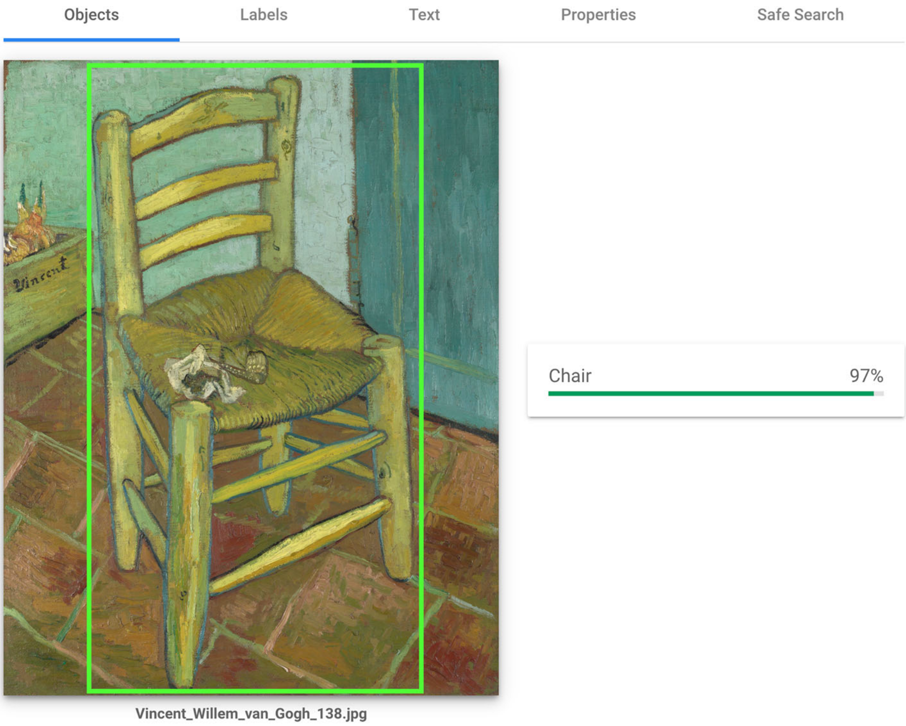
In this essay, we suggest that a "digital" art history is, at the very least, on the horizon, thanks to multimodal machine learning models. These are artificial intelligence (AI) systems that are trained on both textual and visual data; they are what allows a chatbot (or more properly, a large language model) to "see" and interpret images. These multimodal models are capable of far greater nuance and sophistication than their predecessors; they are therefore also biased in much more nuanced and sophisticated ways. They might be said to have a visual culture, or at least a visual ideology. This is the focal point of our analysis. Using multimodal machine learning models, we argue, commits the user to acknowledge, and critically respond to, their respective "machine visual culture." In fact, we argue, multimodal models can *only* contribute to art historical scholarship if this necessary entanglement of analysis and critique is taken seriously. There cannot exist a coherent visual analysis of images using multimodal AI models that is not also, and at the same time, a critique of the conceptual space inherent in the model. Tool and data, in the age of multimodal models, exist in a reciprocal relationship: looking at visual culture through AI means looking at AI through visual culture. Given the current trajectory of AI in academia and elsewhere, this methodological shift is here to stay.

Digital Art History as Object Recognition

In a 2012 interview with Chris Wood, art historian Horst Bredekamp was asked what the impact of machines that can read images will be on the history of art. He answered:

> No, it's impossible for machines. Specialists said recently that even in a thousand years a computer will not be able to recognize the chair painted by Vincent van Gogh as a chair. Computers would need bodies, as the discussion on Körperschema has shown … That is one of the consequences of embodiment philosophy.[12]

As Figure 1 makes clear, Bredekamp was ill-advised. Since 2015, deep convolutional neural networks in particular have led to great leaps in the power of machines to



Chair 97%

Vincent_Willem_van_Gogh_138.jpg

Figure 1. Van Gogh's chair (from *The Bedroom*, 1888) detected by computer vision. We use the Google cloud vision API as an example of a widely used object recognition model.

recognize objects within images. A number of scholars have turned this new capacity for detection onto art-historical problems. In the realm of “thing” detection, the study of iconographic patterns[13] has benefitted from the robustness of pretrained classifiers[14] vis-à-vis figurative works in particular. Pose detection has facilitated the quantitative study of art-historical concepts related to gesture, for instance in Impett’s and Moretti’s study of Aby Warburg’s *Pathosformel*.[15] Finally, face recognition has allowed scholars to study the appearance and reappearance of historical and fictional “characters” at scale.[16]

To be clear: we group all of these approaches[17] together under “object recognition” for two reasons. On the one hand, they all utilize a class of machine learning models – deep convolutional neural networks – that are more or less directly derived from object classification or classification models. On the other hand, and more significantly: all of these approaches attempt to point to (by pixel-wise segmentation, bounding boxes, or classification) semantically unambiguous *things* (“objects”) depicted in images. These models are most often trained on human annotations, and it is assumed that a human observer can verify the model’s predictions – giving some route into identifying, if not explaining, the errors or biases in a model. This form of object recognition (if we are indeed happy to call human limbs and faces “objects”) is thus certainly useful. But what modes of understanding, what forms of knowledge, does it afford? What model of cognition sits underneath an image recognition model?

Vision models for object-related tasks are trained on a predetermined set of object categories: handwritten Arabic numerals in MNIST, or a combination of natural species (bird, cat, marmoset) and manufactured objects (automobile, television, Polaroid camera) in CIFAR-10 or ImageNet. To learn to classify the MNIST dataset, a network does not need to know that “2” falls between “1” and “3”, or indeed that it has anything to do with mathematics or writing. Instead, for neural networks, object/image categories are arbitrary labels that are literally discarded when training, and reattached post hoc, merely for the convenience of the human user. Object detection, in other words, always relies on rigid ontologies which are in themselves meaningless to the computer. The forms of understanding it affords are still necessarily anagraphic, tied to preselected slices of the conceptual, not the physical world. It is this anagraphic quality that also fuels the myth of the objectivity of object recognition. Where there is nothing beyond representation, objectivity, which itself emerges from historical forms of visual representation,[18] becomes the assumed norm.

The ontologies of CIFAR or ImageNet also betray a phantom afterlife of GOFAI (“Good Old-Fashioned AI”), the symbolic, rules-based approach to AI popular until the 1990s. Rules-based systems were good at solving problems like chess, but unable to contribute much to problems of perceiving their environment.[19] As Rodney Brookes put it in *Elephants Don’t Play Chess*: “it is necessary to have its [the AI system’s] representations grounded in the physical world … once this commitment is made, the need for traditional symbolic representations soon fades entirely.”[20] Hubert Dreyfus later famously concluded[21] that GOFAI, even where it took an explicit “Heideggerian” stance such as in the works of Brooks and Philip Agre,[22] fell short of engaging with the world in a truly nonrepresentational way. Historically, then, we could understand object recognition as a simple perceptual appendage onto symbolic AI, even in its “state of the art” manifestation at the beginning of the 2020s.

In digital art history, the essentially anagraphic quality of object recognition sparked attempts to at least improve the ontologies of pretrained systems. "Fine-tuning" such systems, for a while, promised to liberate computer vision from its problematic and anachronistic training datasets, or at least integrate art historical images. Some efforts focused on object recognition in non-photographic images (which computer scientists have called the "cross-depiction problem"[23]), others on recognizing classes of object which were not present in ImageNet;[24] but overall, fine-tuning failed to establish itself as a viable digital methodology for most digital art history projects. The reasons for this are many – they surely include the massive data and compute resources necessary to retrain neural networks. Potential improvements were often insignificant; pretrained models (mostly trained on ImageNet) simply worked well enough for the limited (in the technical, not scholarly sense) tasks designed by digital art history researchers, while the pace of development in "mainstream" computer vision meant that by the time a model had been adapted for art-historical tasks, a better-performing model would have been published elsewhere.

It would also be untrue to say that object recognition models have no sense of the relationship between objects at all. Rather, their relationships are incidental and based on simple visual correlations. This can be seen in the misclassifications made by object recognition systems. Amongst the 10 object classes of the CIFAR-10 dataset, for instance, one algorithm reports especially high confusions between the following pairs of classes: automobile–truck, dog–cat, deer–horse, and bird–airplane.[25] Yet the algorithm has no understanding of what it means to be a vehicle or a pet. These class confusions are entirely visual: birds and airplanes, for instance, are both often photographed against the sky. The object recognition model's internal representations mirror entirely visual relations between objects in the world.

More importantly, however, these class confusions suggest that there might be a more general utility to the internal representations of deep neural networks. Following the hypothesis that a successful classifier must have obtained at least some useful knowledge about the structure of images (in the dataset), "feature extraction" uses selected internal representations to create "embeddings" of images, compressed representations of images as seen by the neural network. Unlike in word-embedding models, where the distributional hypothesis suggests that word correlation has semantic implications at scale,[26] for object recognition models such embeddings remain determined by visual features, and not by any extrinsic understanding of how objects relate to each other in the world. The ability to group together all images containing "horse-like" objects only gets you so far in addressing art-historical questions.

Ultimately, then, the use of embeddings extracted from object recognition models operates within the confines of a labeled visual world. In the context of Drucker's distinction between the digital and the merely digitized, we might consign object recognition to the latter. Though technically impressive, object recognition ultimately serves to facilitate access to images in an automated extension of traditional, categorical metadata.

At the same time, all object recognition models suffer from an implicit vernacular concreteness. It is trivial to observe that the task that object recognition models are designed to solve is valid only in a representational context – where there are no objects, no objects can be detected. But their dependence on the representational goes deeper. "Traditional" computer vision training sets have, since the beginning of the current wave of AI research around 2012, focused almost exclusively on taking (photographic) stock of the vernacular. Intended to train models that would be useful in "real-life" contexts, such training sets contain mostly amateur or stock photographs of everyday objects, scraped from the Web, and labeled through human annotation via platforms like Amazon Mechanical Turk.[27] Rarely would "art" feature in such datasets. Consequently, object recognition models, until very recently, were difficult to adapt to non-vernacular visual culture.[28]

And yet: that it is possible to operationalize notions of similarity that go beyond what would be achievable on the pixel level even with object recognition models points to the significant potential of image embeddings in particular. To fully exploit this potential, however, we need to turn to a more recent class of models in which image embeddings can lead us *beyond* the object recognition paradigm as they function as *deep descriptions of para-visual concepts*: multimodal foundation models.

Multimodal Foundation Models

In multimodal foundation models, we are dealing with two interlocking technical developments. "Foundation model" is a term introduced by researchers at the Stanford HAI institute in 2021. It basically means: models (1) that are very large, and (2) that can be used for a variety of "downstream" tasks. "Multimodality" implies a conjoined processing of text and image data.

In the first instance, their multimodality allows them to encode both text and image data into a common space. They learn not from sets of images with discrete categories, but from pairs of images and texts – digital images with descriptive captions. In practice, this means that their internal representational logic – the shape of their embedding-space – is informed by both visual and linguistic relations. Rather than being constrained by a fixed set of object categories (television, bird, airplane), multimodal models are able to encode an arbitrary textual input (up to a certain length) and images that are well described by that text. These models are implicitly ekphrastic: they learn about texts through images, and vice versa; they necessarily structure both media in relation to the other. This means that they can describe things which are not *objects*, and that they encode both visual and linguistic relations between them; mapping the relation between "airplane," "car," "boat," "travel," "vehicle" in both words and images.

The second distinction is one of scale. Foundation models, which include unimodal text models like GPT-3, are those that are trained on previously inconceivable amounts of data. ImageNet[29] contains around 14 million images in roughly 22,000 categories; at the time of writing, the most common datasets for training text-image models are 400 million and 5 billion (LAION-400M and LAION-5B) image-text pairs respectively. ImageNet is intended to be entirely made up of photographs, ideally with clear, isolated objects at their center; LAION contains drawings, paintings, and born-digital images.

Interestingly, these multimodal datasets have not overcome many of the other issues that "traditional" computer vision datasets were facing, most prominently the issue of bias. In 2019, Trevor Paglen and Kate Crawford[30] exposed the many flaws of ImageNet's human-related categories. Those categories – even if not often used in actual model training, which commonly only used a subset of ImageNet called ILSVRC2012[31] – assembled a host of highly sexist and racist assumptions, such as illustrating the category of "terrorist" almost exclusively with people of Middle Eastern descent. Adam Harvey's more substantial and ongoing work on datasets collected "in the wild"[32] have shown the omnipresence of data collected without consent, and its subsequent, irreversible integration into different machine learning pipelines.

These and many other problems persist – and are amplified – in multimodal datasets. A recent academic critique of a LAION dataset[33] identified "misogyny, pornography, and malignant stereotypes" in LAION-400M. Most recently, child sexual abuse images have been identified in LAION-5B, and it has since been pulled from both academic and industry research. While this represents an important step in addressing the "data" issue of AI, for many multimodal models, the exact composition of the training data is actually unknown. Dataset critique – the dominant mode of critique under the object recognition paradigm – becomes intractable under the multimodal paradigm. AI critique, then, has to move away from the painstaking process of combing through billions of images, and towards analyzing, and critiquing, the model itself – a trajectory that we propose to follow by close-reading a model through a corpus of unrelated images.

Close-reading CLIP

Multimodal foundation models allow for a surprisingly nuanced exploration of complex visual concepts which fall outside the conceptual constraints of object recognition. This affords a new set of practices when working with visual data. At the same time, it also creates new urgency for understanding the "visual culture" of such trained models, precisely because it enables us to look for visual concepts that necessarily only exist in a culturally situated way.

Historically, the paper "Learning Transferable Visual Models from Natural Language Supervision"[34] was a turning point initiating the current paradigm shift. The paper proposes a multimodal foundation model architecture called "CLIP," with both architecture and a pretrained model released on GitHub. CLIP showed impressive capabilities on a wide range of tasks, including "zero-shot" image classification, i.e., the classification of images belonging to classes not specifically observed during training. However, as it was released at the same time as the first DALL-E model, a generative model with impressive image synthesis capabilities and much more discursive impact, CLIP's potential for the digital humanities was not realized until a few months later, when a handful of researchers started to use it as an embedding model in the way described above. CLIP's image embeddings (Figure 2), it turned out, not only surpassed those of previous models in "quality" but also facilitated a completely new form of image retrieval based on its multimodal capabilities. Early multimodal search engines like imgs.ai[35] offered an interface to this functionality.

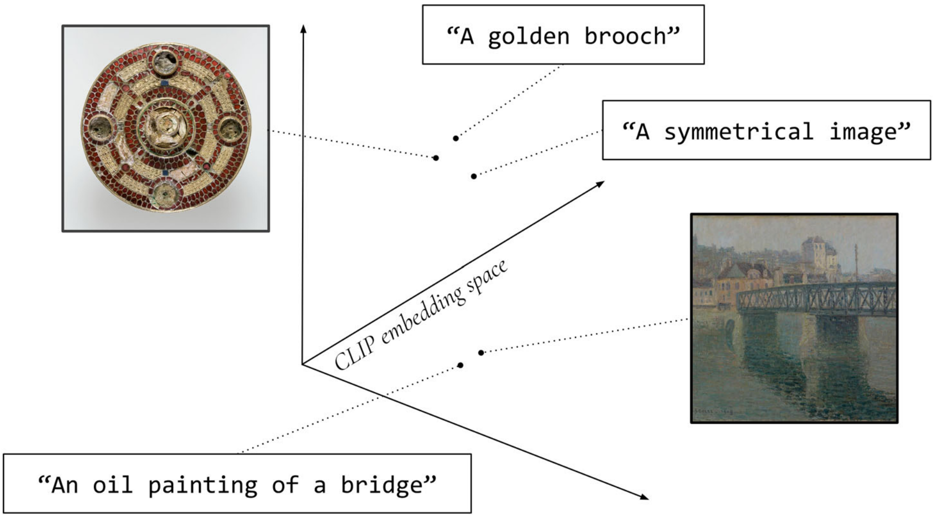


Figure 2. CLIP embeds both texts and images into a common embedding space. The similarity between two texts, between two images, or between an image and a text can thus be estimated by their relative positions in the embedding space. This space is shown as three-dimensional; it is in fact 1024-dimensional. Note that with more dimensions in the embedding space, more complex notions of similarity can be captured; thus, as in the example above, two quite dissimilar texts can be similar to a single image; or, as in Figure 3, the same text ("rhythm") can be similar to a highly heterogeneous set of images.

Through multimodal searching it became possible to implement trivial but difficult to operationalize retrieval tasks like finding images of specific objects, e.g., images of chairs in the Museum of Modern Art, New York, collection, or exploring certain motifs in a collection of magic lantern slides.[36] CLIP, in other words, mirrors the state of the art of object recognition. But more importantly, other than in metadata-based retrieval systems and object recognition models, CLIP, through its natural language interface, allows for retrieval based on *para-visual concepts of arbitrary complexity*. A multimodal network thus has a kind of "mental image" of any text, and indeed a corresponding verbal imaginary of any picture. Recognizing that both these operations stem from the same network and the same training data, we might suggest that rather than mental images or imaginary texts, multimodal networks have *vector imaginary*.

We're not suggesting it's infinitely powerful, of course. Not just because it plays an impossible game in trying to collapse the incommensurable: there is no perfect description of an image, nor an ideal picture of any text. In fact, it is rather poorer at some forms of understanding than one might expect, such as counting or spatial relations ("two cats to the left of a red box") – much to the frustration of the users of image-generation systems, which are themselves often CLIP-based. But whatever the limits of a specific network's understanding, we can at least *ask* for anything we can write down, rather than being limited to choosing from a menu of predetermined – and usually inappropriate – kinds of things (pedestrian, car, cat, dog, airplane …). That does, again, not necessarily mean that the model will always *answer* in the way we might hope; it might just as well fail silently and not adhere to a prompt at all if a token

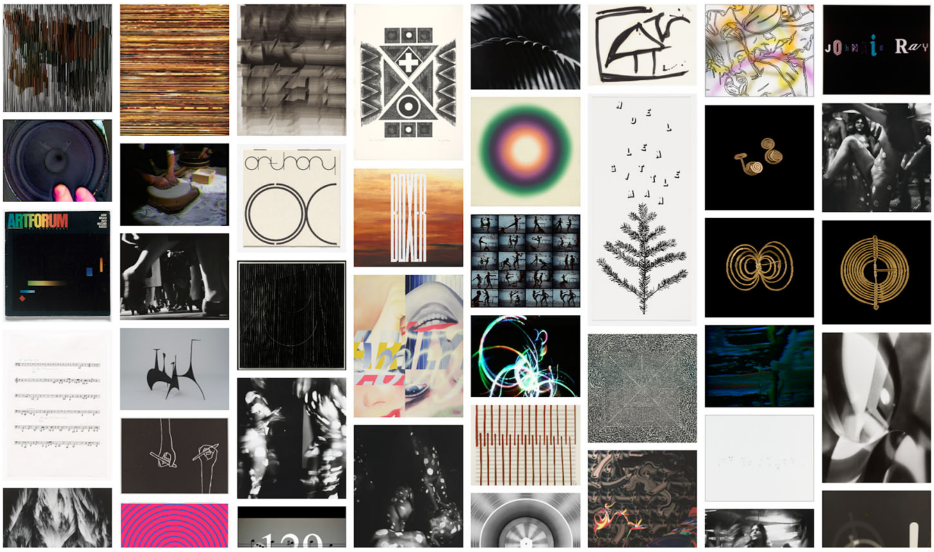


Figure 3. Images for the keyword "rhythm" in the Museum of Modern Art, New York, collection, produced by the imgs.ai search engine. The results show the polyvalence of the knowledge embedded in the CLIP visual artificial intelligence model. Sheet music, album covers, photos of audio equipment, images containing the letters in the word "rhythm," images that could be interpreted as waveforms or spectrograms, as well as "rhythmic" graphical works are returned.

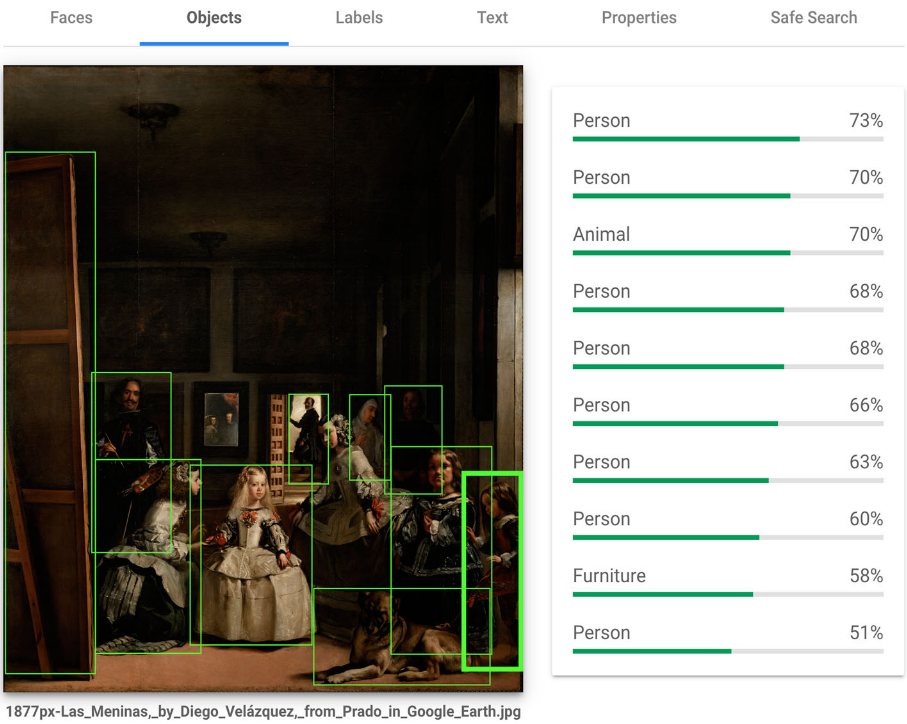
was not well represented enough in the model's dataset. Both the potential and limitations of CLIP, however, are best approximated through concrete examples.

A CLIP search for "summer" will turn up images associated with summer: of beaches, pools, people in swimsuits, or landscapes drenched in yellow light. A CLIP search for "rhythm" (Figure 3) will retrieve images which embody a large spectrum of the meaning inherent in that word: images of sheet music, album covers, and loudspeakers, works that resemble oscilloscope graphs or spectral plots, or graphical works that involve regular patterns that could be described as somehow rhythmic. We are still dealing here with image retrieval, but CLIP takes us far beyond the object recognition paradigm; though perhaps not beyond the logocentric tradition of art history, with its deeply uneven capacity for visual description[37] which the vector logics of neural networks had – for a while – promised to supersede.

Let's consider another example that demonstrates these capabilities of the CLIP model. Diego Velázquez's 1656 painting *Las Meninas* is one of the most discussed pictures of art history. Michel Foucault, famously, spends the whole introduction of *The Order of Things* on it. The painting is famous, in particular, for its play on representation. The painter himself is in the painting, but we do not see what he is painting, as we can only see the backside of the canvas – or do we indeed see what he is painting, as we are potentially looking at the picture he is painting at that moment? There is a mirror which opens up another, invisible image space, and countless gaze relations tell an intricate story about the historical characters in the picture.

Using the techniques of digital art history so far, what can we say about this picture? We might be able to determine the number of people in the picture with the help of a pretrained and/or fine-tuned face detection network. We might confirm the existence of certain image objects – an easel, a dog, other paintings – with the help of an object detection network (Figure 4). We might even be able to estimate the gaze direction of some of the characters in the picture. But under no circumstances could we infer the play on representation that the picture embodies, the fact that it is, with W.J.T. Mitchell, a "metapicture,"[38] a picture about pictures, a representation of representation.

If we run an imgs.ai search for the phrase "Las Meninas" on the collection of the Museum of Modern Art, New York, an institution that does not only not have the famous painting in its collection (which is kept in the Prado in Madrid), but also focuses on contemporary art in general, the results are surprisingly enlightening and show the conceptual depth that CLIP allows the user to access. Among them we find Richard Hamilton's *Picasso's Meninas* from *Homage to Picasso* (1973), which takes up the structure of the Velásquez original but fills it with figures from Picasso paintings. While both works have nothing in common but their compositional structure, CLIP is still able to draw formal connections through their compositional similarity. CLIP also points us to two photographic works, Joel Meyerowitz's *Untitled* from *The French Portfolio* (1980) and Robert Doisneau's *La Dame Indignée* (1948), both explicit plays on representation, mirroring the concern of *Las Meninas* with the gaze relation between people in, and people before,[39] the image.



Person 73%
Person 70%
Animal 70%
Person 68%
Person 68%
Person 66%
Person 63%
Person 60%
Furniture 58%
Person 51%

1877px-Las_Meninas,_by_Diego_Velázquez,_from_Prado_in_Google_Earth.jpg

Figure 4. Google cloud vision API results for *Las Meninas*; the easel is classed as "furniture."

One might ask though if we are not, again, regressing into a representational context here. After all, even if we are reading a work from the Spanish Golden Age *in terms of* a work of contemporary, twentieth-century photography, our reference remains figurative. Is CLIP – in that sense – still bound to the object recognition paradigm we above suggested it transcends? A third, non-exhaustive series of retrieval configurations shows that CLIP, indeed, has learned to "read" abstraction as well.

If we use imgs.ai to search for "Mondrian" in the Rijksmuseum collection, for instance, our results will, unsurprisingly, contain works by Theo van Doesburg, Mondrian's lesser-known contemporary who experimented with similar kinds of abstract workflows. In the even less contemporary collection of the Metropolitan Museum, the same search brings up a host of unrelated artifacts that have one thing in common: they prominently feature a color calibration card.

But even in the absence of *any* discernible shape, CLIP still manages to produce useful results. Continuing with the Metropolitan Museum collection, a search for "Malevich" will prominently feature "empty" black and white pages of books which nevertheless manage to produce a similar focus on texture in the viewer as the "original" black or white works by the Russian suprematist (mediated, of course, by his own internet reception-history, which orbits around the *Black Square*). A search for "Rothko" brings up pieces of Renaissance textile – more specifically velvety silk – where color and texture are interdependent (Figure 5).

The question then is: what are the *potential applications* and *epistemic implications* of a system that has learned to "understand" – or at least operationalize – complex, para-visual concepts *in terms of* visual attributes? How can it, in Drucker's terms, change the way we understand the objects of our inquiry?

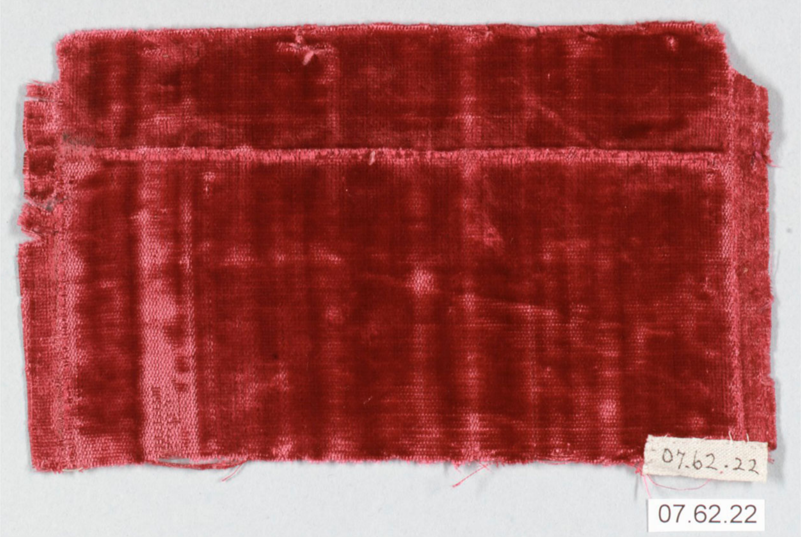


Figure 5. Piece of Italian silk, sixteenth century, Metropolitan Museum, New York. https://www.metmuseum.org/art/collection/search/213060.[64]

To be clear: the quotation marks around "understand" are to be taken seriously. We are not suggesting that CLIP has an "idea" of representation, or abstraction. In that sense, CLIP is not "better" than object detection models, and not closer to any meaningful human-like capabilities. CLIP cannot, for instance, *explain* any of the para-visual concepts it has learned; they can only ever be approximated through examples, they can only ever be read through corpora that are already meaningful. Its approximation of abstraction is just that – an approximation – and thus necessarily less "precise" than any conceptualization of (visual) abstraction that is itself (conceptually) abstract. In other words: CLIP manages, exactly *without* any ability to conceptualize or abstract, to produce results that suggest an ability to conceptualize or abstract. It is this contradiction that we are interested in.

So, what can we know about the para-visual concepts that CLIP operates on? While we can reason about them theoretically, or on the basis of our knowledge about training datasets, we want to suggest that it is precisely the emerging techniques of multimodal digital art history that can be used to investigate the shapes of CLIP's concepts, so to speak: and, by extension, their epistemic and ideological implications. Two experiments – which are to be read as mere suggestions for potential directions – will serve to demonstrate what we see as a necessary entanglement of object and model analysis. We make the tools developed for both experiments freely available on the Web[40] so that our results can be reproduced on demand.

Conceptual Maps: What Makes Paris Look Like Paris?

Again we will consider a concrete example: the city of Paris. CLIP has a concept, a vector imaginary, of Paris. A CLIP-search for "Paris" in the Museum of Modern Art collection, for instance, turns up images of the Tour Eiffel, of Notre Dame, and of generally "French-looking" streets. Paris, for CLIP, is bound up with a set of particular visual, we might say symbolic, properties; but how do they relate to the material reality of Paris as it exists today?[41] Where in Paris looks (to CLIP) like Paris – and what does that tell us about CLIP?

To attempt to answer this question, we collect 10,000 images of Paris from Google Street View, sampled at regular intervals of latitude and longitude. All are images of Paris, the place; and yet not all are equally associated, in the visual logic of CLIP, with the prompt "A photo of Paris." If we plot the strength of the association of each image with that prompt in the CLIP latent space, in the form of a heat map overlay, we can see that CLIP's concept of Paris is highly spatially uneven (Figures 6 and 7). The city center produces the highest associations (especially landmarks like the Arc de Triomphe); as we get to the periphery, the score is significantly lower. CLIP's visual model, its symbolic imaginary, of *Parisness* is one of landmarks and Hausmannian boulevards.

This is an extreme case, of course. And yet it points to a wider sense in which CLIP's model of anything is always a vector imaginary, necessarily deeply culturally situated, linked to the image-economies of the internet: in this case, a visual imaginary of Paris linked both to tourism and to national identity (and state power). Every city (and more broadly, every visual concept) comes with its own highly symbolic

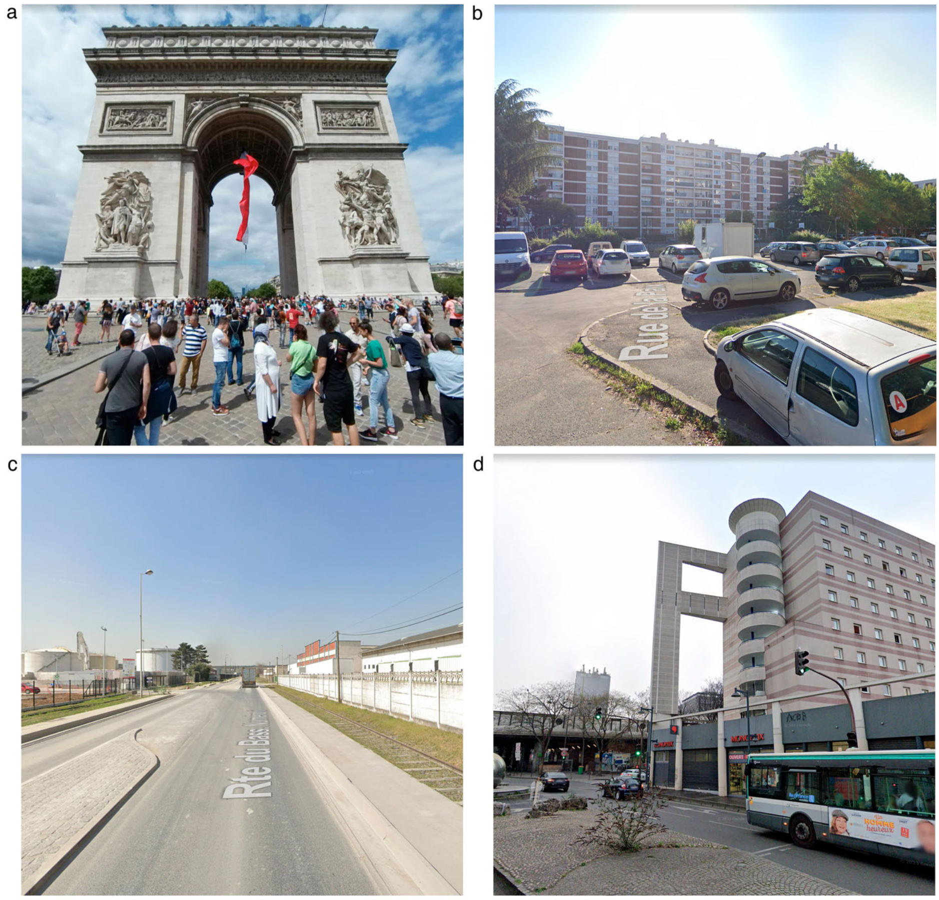


Figure 6. Left to right, top to bottom: two of the most and least Parisian images of Paris; a very "Los Angeles" image of Paris; and the most "New York" image of Paris.

imaginary: the most "Los Angeles" parts of Paris, for CLIP, are to be found in the periphery, not in the center. One of the most "dystopian" parts of Paris is, at the same time, its most "New York" part – the actually most "dystopian" image being an inside shot of a laser tag facility somewhere on the outskirts of the city.

Gardens, graffiti, public spaces, gentrification: one can imagine using multimodal networks such as CLIP in this way to literally *map out* urban cultural geography. In proposing this experiment, we are not calling for multimodal AI to become a new tool in the arsenal of architects or urban planners. As discussed above, the training data of large multimodal models such as CLIP are enormous and wide-ranging, but full of the biases of the wider internet – in race, gender, class, or cultural Eurocentricity.

What we are proposing is rather more like the inverse. Visual studies (or architecture, urbanism, psychogeography) can be turned to the urgent task of understanding the visual ideology of multimodal AI, precisely by mapping the network's culturally situated mental images of textual terms. This holds, of course, not just for street

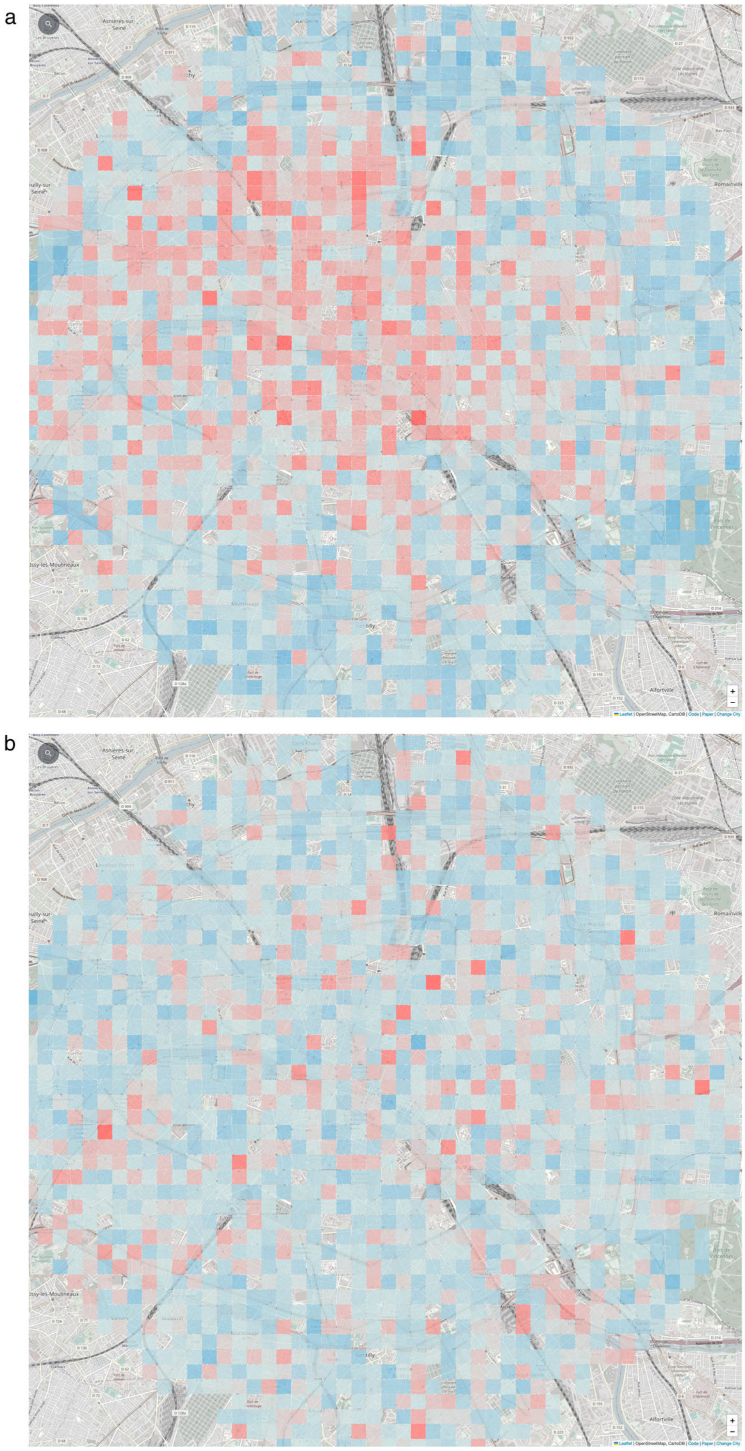

a
Asnières-sur-Seine
Issy-les-Moulineaux
Romainville
Alfortville
Leaflet | OpenStreetMap, CartoDB | Code | Paper | Change City
b
Asnières-sur-Seine
Issy-les-Moulineaux
Romainville
Alfortville
Leaflet | OpenStreetMap, CartoDB | Code | Paper | Change City

Figure 7. Strength of CLIP activations for sampled Google Street View images for "a photo of Paris" and "a photo of Los Angeles"; red is a stronger association, blue is weaker.

view, but for any image corpus. Using CLIP as a research tool to study visual culture is always, in fact, a way of using visual culture to explore CLIP's own way of seeing.

2D CLIP: Naked and Nude

Let us pivot our attention to the more common object of study in digital art history: the corpus of paintings. Multimodal networks give us a new way of measuring visual phenomena: we can measure anything we can name across a dataset of images. And though we should not expect to measure it accurately or fairly, it is precisely the distortions produced by CLIP's incomplete or unexpected interpretations of a textual concept that give us a way into exploring its latent visual culture.

Instead of using these text-image similarities as rankings in a search engine or colors on a map, we extract them as "features." Building on the work of Lev Manovich (amongst others), each feature can then become an axis on which to build a visualization of an image corpus.

One of the most powerful ways of probing (and indeed using) CLIP's vector imaginaries is to focus on the messy distinctions between a pair of highly entangled concepts; as between "Paris" the word, and "Paris" the set of street-view images. This is the intuition behind our browser-based visualization tool, *2D CLIP*.[42] In 2D CLIP, two separate CLIP-based visual concepts are measured *against each other*. We are, here, attempting to use one disentanglement, between tightly overlapping concepts, to work through another: between the AI model and the art-historical experiment; between the training data and the data for analysis; between lens and object.

Entangled conceptual pairings are often themselves the focus of critical debate. Critics often seem to learn something by working through this kind of desynonymization: interdisciplinary/transdisciplinary, modernism/modernity, Kunstgeschichte/ Bildwissenschaft.[43] In one of its forms – *différance* – this "systematic play of differences"[44] became a main weapon of Derridarean deconstruction.

"The naked, therefore, who compete / Against the nude may know defeat," wrote Robert Graves in 1957.[45] A year earlier, Kenneth Clark had published an influential book on the art-historical nude which began with a demarcation from the merely *naked*.[46] This distinction does not appear in most European languages, making this a rather Anglophone debate — Clark is thus indebted to the "elaborate generosity" of English, though he admits that nude (as a noun, at least) was only "forced into our vocabulary" by critics of the eighteenth century.[47] Nakedness, for Clark, implied embarrassment; nudity, on the other hand, was a "balanced, prosperous and confident body."

It was in direct response to Clark's views on the nude that John Berger coined the term *male gaze*: a body, for Berger, can be naked on its own, but it "has to be seen as an object in order to become a nude."[48] For Berger, who spends less time than Clark discussing male nudes, the distinction is one of commodification rather than confidence.

The two terms even appear in a very early, radical example of quantitative art history. "Less than 5% of the artists in the Modern Art sections are women, but 85% of the nudes are female," reads the Guerilla Girls' famous poster[49] "Do women have to be naked to get into the Met. Museum?"

Is the distinction one of sexualization, of commodification, of status as art-object? For Linda Nead, it relates not only to the boundaries of pornography, but to "mind/body, form/matter, art/obscenity … the beautiful and the sublime." Ultimately, these lead to the "distinctions between inside and outside, between finite form and form without limit."[50]

Naked/nude, then, is a disjuncture that points outwards, to encompass a whole set of categorical tensions within art history. The perfect duplet, in other words, to start exploring the ideology CLIP's vector imaginary. We have taken 1000 images from the collection of the Art Institute of Chicago, and mapped them along their CLIP-similarities with the terms "naked" and "nude" (Figure 8).

As we would expect for such tightly enmeshed concepts, the images from the Art Institute follow a fairly straight line: images which are more naked, are also more nude. But there is some noise, some deviation to the pattern. Most images do not sit on the imaginary diagonal line, but just above (more nude) or below (more

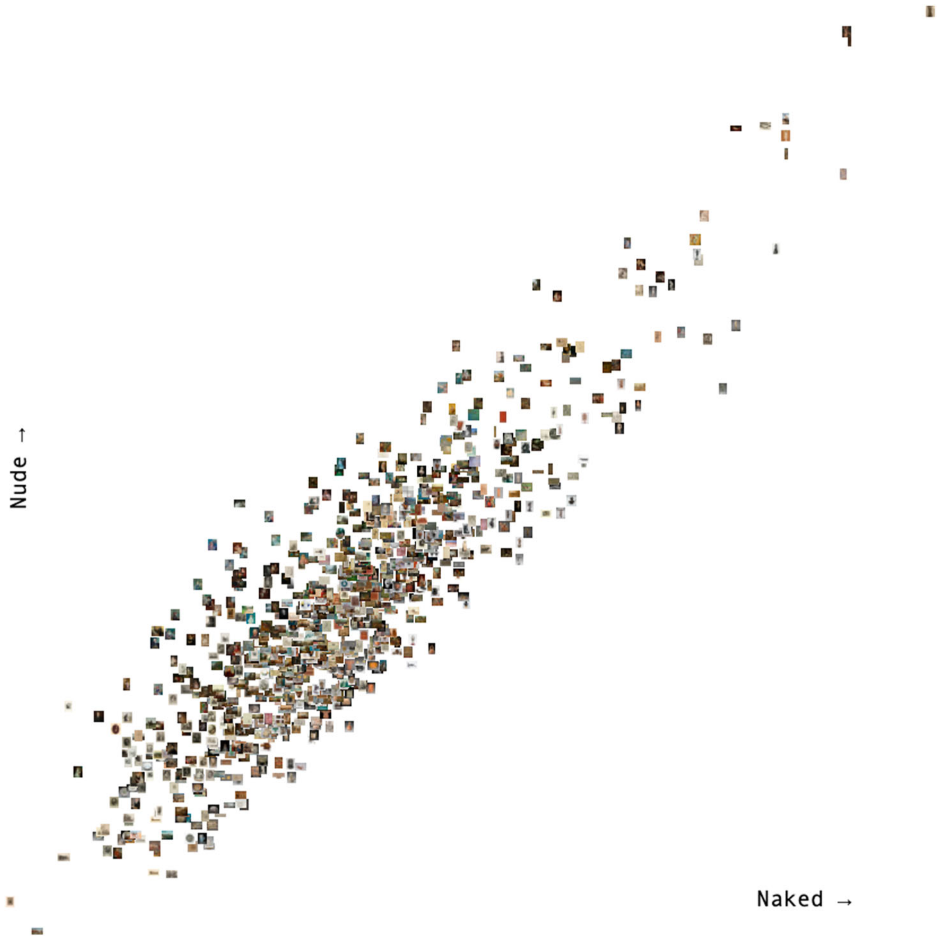


Figure 8. 2D CLIP interface with 1000 images from the Art Institute of Chicago collection scattered along two conceptual axes: "naked" and "nude."

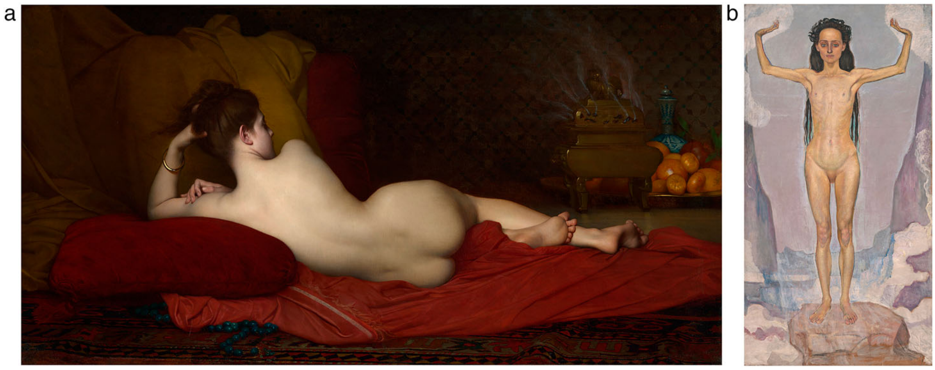


Figure 9. CLIP's "Nude" and "Naked" – Jules Joseph Lefebvre's *Odalisque* (1874) and Ferdinand Hodler's *Day (Truth)* (1896/1898).

naked). For CLIP, Jules Joseph Lefebvre's late nineteenth-century reclining *Odalisque* (literally "chambermaid," but in the orientalist imaginary, closer to "concubine") is markedly more nude, as are a host of reclining and seated unclothed women.

CLIP's nudes largely share this stillness, but there are exceptions: the sexual violence of Tintoretto's *Tarquin and Lucretia*, for instance. Ferdinand Hodler's *Day (Truth)* is, by contrast, determinedly *naked* for CLIP (Figure 9). How are we to read that against Clark's *balanced, prosperous, confident*?

This exercise does not just tell us about CLIP, but about the Art Institute's own collection (or our subsample of it). Most painted nudity is female nudity, of course. Unclothed male bodies are either to be found either together with female ones (Adam and Eve, Tarquin and Lucretia), or in sculpture. CLIP sees these largely classical male sculptures as decidedly *naked*. Strange, given their central place in the imaginary of the nude: they include a Roman marble Meleager and a bronze Hercules after Lysippos, whom Clark had said created "some of the finest nudes in art."[51] Though rendered somewhat androgynous by its missing head and limbs, even the Art Institute's Knidos Aphrodite – a copy of Praxiteles, no less – is seen by CLIP as less nude than naked.

CLIP's nude seems strongly gendered; but might it also be dependent on medium? And CLIP's privileged medium – following the logic of computer vision – is photography, with Alfred Stieglitz's photograph of Georgia O'Keeffe's bare torso being seen by CLIP as both more naked, and more nude, than the rest of the collection: we find it at the extreme top-right of the distribution. The most naked image is the most nude: we might return to Lynda Nead, who insists that the two are always inseparably entangled, that there "can be no naked 'other' to the nude, for the body is always already in representation."[52] This mirror, we suggest, our new methodological and epistemic entanglement: in the age of foundation models, we can never quite isolate the neural network from the object of study.

We have merely sketched here a proposal for a new way of working, and the operationalized naked–nude conceptual entanglement itself requires much more interpretative work. How typical are the results above; how much have we learned about the collection of the Art Institute, in other words, and how much about CLIP? Could the gendered logic of CLIP be investigated not just multimodally, but also through text–text and image–image similarities? How has the gendering of other words influenced CLIP's understanding of these two terms? Are CLIP's visual projections of the two concepts in paintings similar to those in photographs? What is the role of advertising, of images from social media, of pornography, in the shape of CLIP's embedding-space;[53] and without access to the original training dataset, how might we tell?

Conclusion

Our experiments with using CLIP for image retrieval have illuminated how its embeddings take us far beyond the object recognition paradigm. While it is safe to assume that CLIP is limited in many of the same ways that other foundation models are, the breadth and complexity of its para-visual concepts has yet to be fully exploited in digital art history and digital visual studies. Our case studies – using CLIP to illuminate the urban fabric of Paris, while at the same time illuminating the sociogeographic biases of the model; using CLIP to operationalize the naked–nude distinction, while at the same time exposing the model's gendered visual logic – have no more than hinted at potential areas of application and critique. They are demonstrations, first and foremost, of what still cannot be operationalized.

We do not claim, then, to have produced one of the "dramatic proof-of-concept works"[54] that Drucker sees necessary to advance from digitized to digital art history, or that "improving" CLIP would fully align it with art history. On the contrary: that CLIP produces meaningful approximations without a way of reasoning, without explicitly and consciously moving from patterns to concepts, if anything, positions it further away from traditional ways of "doing" art history. Not in terms of its "machine visual culture" which emerges from the further (digital) canonization of the canon – but in terms of its "machine methodology." With CLIP, we can only ever investigate *this* particular slice of the canon, using *this* idiosyncratic notion of similarity and dissimilarity. Object and method are inextricably linked.

And yet, the fast adoption of CLIP-based retrieval methods in all areas of cultural heritage work,[55] pragmatically, suggests that multimodal foundation models are here to stay, and with them some of the methods described above. We propose to investigate CLIP (as a metonym for multimodal foundation models more generally) precisely because, for better or worse, they will be widely used in the years to come.

Before the age of foundation models, neural networks never had an interesting model of human visual culture. The many deficiencies of ImageNet and its cousins were so easily exposed because of its claim to universality, a claim which was already made for the simple conceptual ontology ImageNet builds upon.[56] Today, the makers of Stable Diffusion announce that their model "is the culmination of many hours of collective effort to create a single file that compresses the visual information of humanity into a few gigabytes."[57] This is still a ridiculous thing to say, of course: digitized images are nothing but the tip of the vast iceberg of human image production.

And yet, while contemporary visual models are often still not good enough to facilitate those hard technical challenges that AI researchers are actually aiming for (like detecting cancer or driving a car), they contain so much more information about human visual culture[58] that digital art history might be the one discipline where they actually will have an outsized impact. But their usefulness – this is the main argument we would like to make with this text – hinges on a reevaluation of what digital art history actually is as a field. As Ted Underwood writes: "approaching neural models as models of culture rather than intelligence gives us more reason to worry about them. But it also gives us more reason to hope."[59]

To put it bluntly: we wanted digital art history, and we got digital art history, but it is not what we expected. If we want to go beyond the object recognition paradigm, if we want to go beyond merely taking stock of images and image objects, if we want to integrate close and distant viewing, if, in other words, we want to move towards a "digital" art history, we have to accept that the scope of the field needs to expand. Models – and their idiosyncratic ways of seeing the world – are our responsibility now, and any art-historical study harnessing the power of contemporary machine learning must necessarily, at least in part, also be a study *of* contemporary machine learning.[60]

To translate this into disciplinary terms: digital art history needs to open up to influences, and cross-collaborations from two directions. The first is from media studies. If we truly cannot have our cake and eat it too – if we cannot harvest the utility of foundation models without them becoming our object of study – we need a critical apparatus that is up to the task, a way to talk about visual models and models of the visual. In particular, recent work in critical AI studies[61] – a field which itself is only forming now – suggests that this methodological revision is already on the way in other disciplines.

The second direction is from computational literary studies. Scholars like Ted Underwood, Rita Raley, or Matt Kirschenbaum have been considering the impact of large language models on the study of literature since the early days of the BERT model. Humanist critique has always emphasized the split between the (formal) language of computers and their perceived affordances, e.g., an ideologically discrete view of the world where everything is a binary distinction, and the potential for ambiguity and polysemy in natural languages. In large-scale vision models, however, the medium of formal inference is now natural language – or, at the very least, a vector space which appears to be equally continuous, ambiguous, and polysemic.

Art history is used to negotiating the mediation of the verbal; but we now triangulate between the double mediation of text and code. This complicates the critical analysis of machine learning systems in general, but also presents an opportunity to consolidate humanist and technical critique. "Prompt engineering" means employing the complex, multilayered features of language that humanists already have at their disposal as tools of visual analysis.

The discipline of art history, we conclude, is well prepared for this shift. In fact, we would like to suggest that art history is uniquely well prepared for what is to come. In 2013, Drucker suggested that "*the crucial recognition that digitization is not representation but interpretation* will serve as a critical springboard for insight."[62] And indeed, that digital corpora "carry interpretative inflection"[63] could be described as a disciplinary consensus in 2023. But it turns out that Drucker's critique applies to those advanced computational methods that would constitute a truly "digital" art history as well, and it does so in an even stronger sense. The advent of multimodal foundation models forces us to acknowledge that their critique is not just a prerequisite, but indispensable to their application. Rather than describing the biases, glitches, limitations, and anachronisms of machine vision as limitations of digital art history, they are reconstituted as its object of study – as phenomena which art historians are well-equipped to explore.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

LEONARDO IMPETT, PhD (EPFL, 2020), is a University Assistant Professor in Digital Humanities at the University of Cambridge, UK. His research sits between visual culture and computer vision. He was previously Assistant Professor of Computer Science at Durham University, and has a background in information engineering and machine learning. He has previously been based at EPFL, Villa I Tatti (the Harvard University Center for Italian Renaissance Studies), and the Bibliotheca Hertziana (Max Planck Institute for Art History in Rome), and has been an investigator on projects funded by the Arts and Humanities Research Council and the Volkswagen Foundation. Website: https://leoimpett.github.io

FABIAN OFFERT, PhD (UCSB, 2020), is Assistant Professor for the History and Theory of the Digital Humanities at the University of California, Santa Barbara, USA. His research and teaching focuses on the visual digital humanities, with a special interest in the epistemology and aesthetics of computer vision and machine learning. He is principal investigator of the international research project "AI Forensics" (2022–2025), funded by the Volkswagen Foundation, and was principal investigator of the UCHRI multicampus research group "Critical Machine Learning Studies" (2021–2022). Before joining the faculty at UCSB, he served as postdoctoral researcher in the German Research Foundation's special interest group "The Digital Image," associated researcher in the Critical Artificial Intelligence Group (KIM) at Karlsruhe University of Arts and Design, and Assistant Curator at ZKM Karlsruhe, Germany. Website: https://zentralwerkstatt.org

ORCID

Leonardo Impett http://orcid.org/0000-0003-1774-5175

Notes

[1] Johanna Drucker, "Is There a 'Digital' Art History?," *Visual Resources* 29, no. 1–2 (2013): 5–14; p. 6. Claire Bishop, five years later, doubles down on Drucker's claims: "Theoretical problems are steamrollered flat by the weight of data." Claire Bishop, "Against Digital Art History," *International Journal for Digital Art History* 3 (2018): 125.

[2] Drucker, "Is There a 'Digital' Art History?," 7.

[3] Drucker, "Is There a 'Digital' Art History?," 5–6.

[4] The recurring accusation that the digital humanities are a "neoliberal" field, or at least complicit in the radical transformation of the university under late capitalism, hinges on this foundational search for the "surplus value" of the digital. Traditional humanistic methods are already saying everything there is to say about culture – or so the

conservative argument goes – such that any additional mode of analysis becomes extractive, i.e., destructive.

[5] Franco Moretti, *Distant Reading* (London: Verso Books, 2013).

[6] Taylor Arnold and Lauren Tilton, "Distant Viewing: Analyzing Large Visual Corpora," *Digital Scholarship in the Humanities* 34, suppl. 1 (2019): i3–i16.

[7] Diana Seave Greenwald, *Painting by Numbers: Data-driven Histories of Nineteenth-century Art* (Princeton, NJ: Princeton University Press, 2021).

[8] Matthew David Lincoln, "Modeling the Network of Dutch and Flemish Print Production, 1550–1750" (PhD diss., University of Maryland, College Park, 2016).

[9] Even bigger challenges emerge from abstract art, sculpture, or performance art. For the latter, see Miguel Escobar Varela, *Theater as Data: Computational Journeys into Theater Research* (University of Michigan Press, 2020).

[10] Lev Manovich, "Style Space: How to Compare Image Sets and Follow Their Evolution," 2011, http://manovich.net/content/04-projects/073-style-space/70_article_2011.pdf (accessed July 15, 2023). See also Lev Manovich, *Cultural Analytics* (Cambridge, MA: MIT Press, 2020).

[11] For a history of computer vision that touches upon some of these differences, see James Dobson, *The Birth of Computer Vision* (University of Minnesota Press, 2023).

[12] Christopher Wood and Horst Bredekamp, "Iconoclasts and Iconophiles: Horst Bredekamp in Conversation with Christopher Wood," *Art Bulletin*, December 2012: 515–27; p. 526f.

[13] Peter Bell and Leonardo Impett, "The Choreography of the Annunciation through a Computational Eye," *Histoire de l'Art* 34, no. 87 (2021): 1–6.

[14] Artificial neural networks trained on standard datasets, most prominently ImageNet.

[15] Leonardo Impett and Franco Moretti, "Totentanz. Operationalizing Aby Warburg's *Pathosformeln*," Stanford Literary Lab Pamphlet 16, 2017. See also Peter Bell and Leonardo Impett, "Ikonographie und Interaktion. Computergestützte Analyse von Posen in Bildern der Heilsgeschichte," *Das Mittelalter* 24, no. 1 (2019): 31–53.

[16] Prathmesh Madhu, et al., "Recognizing Characters in Art History Using Deep Learning," *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia Heritage Contents* (New York, NY: Association for Computing Machinery, 2019): 15–22.

[17] A recent overview and critique of the state of the art of object recognition can be found in Amanda Wasielewski, *Computational Formalism. Art History and Machine Learning* (Cambridge, MA: MIT Press, 2023).

[18] Lorraine Daston and Peter Galison, *Objectivity* (Princeton, NJ: Princeton University Press, 2021).

[19] Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Cambridge, MA: Harvard University Press, 1988).

[20] Rodney A. Brooks, "Elephants Don't Play Chess," *Robotics and Autonomous Systems* 6, no. 1–2 (1990): 3–15; p. 5.

[21] Hubert L. Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," *Philosophical Psychology* 20, no. 2 (2007): 247–68.

[22] Philip E. Agre, *Computation and Human Experience* (Cambridge, UK: Cambridge University Press, 1997).

[23] Peter Hall, et al., "Cross-depiction Problem: Recognition and Synthesis of Photographs and Artwork," *Computational Visual Media* 1 (2015): 91–103.

24 c.f. the *Saint George on a Bike* project of the Barcelona Supercomputing Center and Europeana, URL: https://saintgeorgeonabike.eu/

25 Qun Liu and Supratik Mukhopadhyay, "Unsupervised learning using pretrained CNN and associative memory bank," *International Joint Conference on Neural Networks* (New York, NY: IEEE, 2018).

26 The classic example is the emergence of rudimentary analogic reasoning capabilities in word-embedding models, such that analogy questions like "what is to woman what king is to man?" can be answered by simple vector mathematics ("king + woman – man").

27 The labor issues surrounding both object recognition and multimodal models (e.g., the globalization of clickwork) are beyond the scope of this article. For a discussion, see for instance Lilly C. Irani and M. Six Silberman, "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery, 2013): 611–20.

28 For example, Prathmesh Madhu, et al., "Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning," *ACM Journal on Computing and Cultural Heritage* 16, no. 1 (2022): 1–17.

29 ImageNet (in its reduced ILSVRC form) was the most popular research dataset of the pre-multimodal age and facilitated many early deep learning breakthroughs. See for instance Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25 (2012).

30 Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Images in Machine Learning Training Sets," *AI & Society* 36, no. 4 (2021): 1105–16.

31 Olga Russakovsky, et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115 (2015): 211–52.

32 Adam Harvey and Jules LaPlace, "Exposing.ai" (2021), https://exposing.ai/ (accessed January 15, 2023).

33 Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes," arXiv preprint 2110.01963 (2021).

34 Alec Radford, et al., "Learning Transferable Visual Models from Natural Language Supervision," *International Conference on Machine Learning (ICML)* (Brookline, MA: Microtome Publishing, 2021): 8748–63.

35 Fabian Offert and Peter Bell, "imgs.ai. A Deep Visual Search Engine for Digital Art History," *International Journal for Digital Art History* (forthcoming).

36 Thomas Smits and Melvin Wevers, "A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections," *Digital Scholarship in the Humanities* 38, no. 3 (2023): 1267–1280.

37 This unevenness – between, say, the description of Euclidean geometry and that of textures – is tackled in Michael Baxandall, "The Language of Art History," *New Literary History* 10, no. 3 (1979): 453–65.

38 W. J. Thomas Mitchell, *Picture Theory: Essays on Verbal and Visual Representation* (Chicago, IL: University of Chicago Press, 1995), 35.

39 Using George Didi-Huberman's term. See Georges Didi-Huberman, *Confronting Images: Questioning the Ends of a Certain History of Art* (University Park, PA: Penn State Press, 2005).

40 https://leoimpett.github.io/clip-map/, https://leoimpett.github.io/2dclip/ _

41 The inspiration for this experiment comes from a now classic computer science paper that aimed to isolate geographically salient architectural elements ("windows, balconies and street signs"), taking Paris as an example. See Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros, "What Makes Paris Look Like Paris?," *ACM Transactions on Graphics* 31, no. 4 (2021): 1–9.

42 The tool runs client-side in the browser, and is thus independent of network limitations and server-side restrictions. It allows researchers to load their own, locally hosted image collections into an accessible interface to create their own conceptual scatter plots.

43 "Art history, image-science" – although the latter term is notoriously hard to translate (as indeed are both its constituent roots).

44 Jacques Derrida, "Différance" (trans. Alan Bass), in *Margins of Philosophy* (Chicago, IL: University of Chicago Press, 1982): 3-27.

45 Robert Graves' 1957 poem *The Naked And The Nude.* See Eugene Hollahan, "Sir Kenneth Clark's The Nude: Catalyst for Robert Graves's 'The Naked and the Nude'?", *PMLA* 87.3 (1972): 443–51.

46 Kenneth Clark, *The Nude. A Study in Ideal Form* (New York, NY: Pantheon Books, 1956); p. 23; in fact the distinction was probably made in the early 1950s, when the lectures that formed the basis of the book were delivered.

47 We might note that it carries uncomfortable echoes of the common English distinctions between (lower) Germanic and (higher) French terms for meat: cow/beef, calf/veal, sheep/mutton.

48 John Berger, *Ways of Seeing* (London: Penguin Books, 1972): 54.

49 Guerrilla Girls, *Do Women Have to Be Naked to Get Into the Met. Museum?*, 1989 [poster]. www.guerrillagirls.com.

50 Lynda Nead, *The Female Nude* (Abingdon: Routledge, 2001): 22, 11.

51 Kenneth Clark, *The Nude*, 181.

52 Lynda Nead, *The Female Nude*, 16.

53 Recently, David Thiel has found hundreds of images of suspected child sexual abuse in LAION-5B, a dataset not used to train CLIP, but used to train a number of image-generating AI systems. See David Thiel, *Identifying and Eliminating CSAM in Generative ML Training Data and Models*, Technical Report, Stanford University, 2023.

54 Johanna Drucker, "Is There a 'Digital' Art History?," 12.

55 See for instance Matthias Springstein, et al., "iART: A Search Engine for Art-Historical Images to Support Research in the Humanities," *Proceedings of the 29th ACM International Conference on Multimedia*] (New York, NY: Association for Computing Machinery, 2021): 2801–3; and Marcos V. Conde and Kerem Turgutlu, "CLIP-Art: Contrastive Pre-training for Fine-grained Art Classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE, 2021): 3956–60.

56 ImageNet builds upon the ontology of WordNet, see George A. Miller, et al., "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography* 3, no. 4 (1990): 235–44.

57 Emad Mostaque, "Stable Diffusion Public Release," https://stability.ai/blog/stable-diffusion-public-release (accessed July 15, 2023).

58 The information learned by such models has, of course, still to be contextualized in relation to potentially significant dataset and inductive biases. For the former, see

our discussion above. For the latter, see Eva Cetinic, "The Myth of Culturally Agnostic AI Models," arXiv preprint 2211.15271 (2022), and Fabian Offert, "On the Concept of History (in Foundation Models)," *IMAGE* 37, (2023): 121–134. As Ted Underwood writes: "approaching neural models as models of culture rather than intelligence gives us more reason to worry about them. But it also gives us more reason to hope." Ted Underwood, "Mapping the Latent Spaces of Culture," *Startwords* 3, August 2022, https://startwords.cdh.princeton.edu/issues/3/mapping-latent-spaces/ (accessed July 15, 2023).

[59] Ted Underwood, "Mapping the Latent Spaces of Culture,". See also Sonja Drimmer's and Christopher J. Nygren's recent "axioms" on art history and artificial intelligence, where they caution against an uncritical adoption of artificial intelligence in art history in general, and digital art history in particular: Sonja Drimmer and Christopher J. Nygren, "Art History and AI: Ten Axioms," *International Journal for Digital Art History* 10 (2023)5:03-5:13. https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/90400/89769

[60] Fabian Offert and Peter Bell, "imgs.ai. A Deep Visual Search Engine for Digital Art History," *International Journal for Digital Art History* (forthcoming); See also David Berry, "The Explainability Turn," *DHQ: Digital Humanities Quarterly* 17, no. 2 (2023). https://web.archive.org/web/20231211003002/https://dhq-static.digitalhumanities.org/pdf/000685.pdf

[61] See for instance the recent *American Literature* special issue on "Critical AI: A Field in Formation," edited by Rita Raley and Jennifer Rhee (*American Literature* 95, no. 2, 2023).

[62] Johanna Drucker, "Is There a 'Digital' Art History?," 12, original emphasis.

[63] Johanna Drucker, "Is There a 'Digital' Art History?," 12.

[64] The URL is provided here as work has very little metadata and might not be findable.