COMMUNICATION & SOCIETY

Miscellaneous

Visual Motifs and Artificial Intelligence: Developing Machine Learning Models Based on Comparative Iconography

Adam Phillips
https://orcid.org/0009-0007-3329-389X
adam.phillips@upf.edu
Universitat Pompeu Fabra

Daniel Grandes Rodriguez
https://orcid.org/0009-0001-3767-2471
daniel.grandes01@estudiant.upf.edu
Universitat Pompeu Fabra

Miriam Sánchez-Manzano
https://orcid.org/0000-0002-4103-9118
miriam.sanchez@upf.edu
Universitat Pompeu Fabra

Alan Salvadó
https://orcid.org/0000-0001-8282-2021
alan.salvado@upf.edu
Universitat Pompeu Fabra

Submitted
July 6th, 2024
Approved
April 1st, 2025

© 2025
Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
www.communication-society.com

2025 – Vol. 38 (2)
pp. 218-235

How to cite this article:
Phillips, A., Grandes Rodriguez, D., Sánchez-Manzano, M., & Salvadó, A. (2025). Visual Motifs and Artificial Intelligence: Developing Machine Learning Models Based on Comparative Iconography, *Communication & Society*, *38*(2) 218-235.
https://doi.org/10.15581/003.38.2.016

Abstract

Can new AI datasets go beyond the hierarchical logics of imitation and replication, relating images from different media, comparatively? This article tries to answer this question, and poses new ones, by sharing the methodological foundations and the preliminary results of a research project we are currently developing at Pompeu Fabra University, entitled *Visual Motifs Identification and Comparative Image Learning*. The basis of this project is to combine the machine learning and computer vision background of mathematicians and engineers with the humanistic expertise of art and film historians, in order to foster a radically different approach to AI datasets and models. Instead of developing algorithms capable of imitating or replicating styles and artists, we propose a new working methodology rooted in the concept of visual motif. What visual motifs offer, compared with existing models that use computer vision strategies, is a more nuanced and refined interpretation of images, based not only on standard recognition of geometrical or semantic data but on the meaningful aesthetic and ideological choices of previous creators through art and media history. Instead of isolating images in a given medium, genre or period, the aim of visual motifs is to juxtapose, compare and discriminate across multiple artforms (painting, sculpture, cinema, photography, video games, comic books, etc.). Therefore, we are first curating a dataset using this comparative methodology, and then training a machine learning model capable of recognizing different visual motifs in a previously unseen image, with an aesthetic and critical background.

Keywords

Artificial Intelligence, Iconography, Visual Motif, Dataset, Algorithm, Art

Collaboration

Manuel Garin, Gloria Haro and Coloma Ballester are part of our research project, and contributed to the elaboration of the article. Related national projects: PID2021-127643NB-I00 and PID2020-116277GA-I00. Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MICIU/AEI /10.13039/501100011033.

1. Introduction

When it comes to how Artificial Intelligence (AI) models deal with visual art, the focus is often put on the technical perfection in emulating or the stylistic reliability offered by a given algorithm, that is, on the way technology imitates previous works and authors by replicating them. As Joanna Zylinska explains in her book on the subject, when it deals with art, artificial intelligence is "first and foremost a sophisticated agent of pattern recognition" (Zylinska, 2020, p. 25), aimed at reproducing the *what-looks-like* of art. Under that light, and taking into account the ethical debates about how images are generated and used, current technologies are very good at imitating and enhancing visual content: the degree of stylistic accuracy they can achieve is remarkable, sometimes mesmerizing. It may be the unique style of a renowned painter, as in the case of Van Gogh analyzed by Lev Manovich (2020), or the specificity of an artistic movement (like the AI versions of Japanese prints *ukiyo-e*), but beyond each example, the generative models used now by millions of users, such as Stable Diffusion, have improved a lot by replicating aesthetic paradigms, by copying *what-looks-like.* Due to the automated scraping of millions of images from websites (2.3 billion in the case of LAION-2B-EN, the set used for the training of Stable Diffusion (Baio, 2022)), generative AI is capable of creating images *à la manière de* very easily, thus reinforcing the hegemony of canonical names in art history. But what about the ability to think *in between* artists and styles, to relate *what-doesn't-look-like*? What happens when, instead of asking algorithms to copy, we ask them to juxtapose and rethink images from varied artforms comparatively? Can new, critical datasets, go beyond the hierarchal logics of imitation and replication?

This article tries to humbly answer those questions, and raise new ones, by sharing the methodological foundations and the initial results of a research project we are currently developing at Pompeu Fabra University. Entitled *Visual Motifs Identification and Comparative Image Learning*, the very basis of this project is to combine the machine learning and computer vision background of engineers and mathematicians with the humanistic expertise of art historians, in order to foster a radically different approach to AI datasets and models. Its main goal is to develop technology able to automatically identify—and learn from—visual motifs (briefly defined as identifiable compositions and patterns that image-makers use to express things visually), enhancing the capacity of machine learning and automatic visual analysis with the complex models of iconography. What visual motifs offer, compared with existing models that use computer vision strategies, is a more nuanced and refined interpretation of images, based not only on standard recognition of geometrical or semantic data but on the meaningful aesthetic and ideological choices of previous creators through art and media history. Besides, instead of basing the technology on the imitation of previous artistic movements or authors (like the aforementioned examples), visual motifs serve to contrast and relate multiple image typologies and authors, from painting, sculpture or architecture to cinema and photography, in a comparative fashion.

Before digging deeper into the project, it is key to underline that, while its main goal is to develop new machine learning models, our entire approach is rooted in an ethical and philosophical standpoint. That is, the sincere belief that the extractive and replicative logic that has arguably been feeding AI development for the last decade (more, faster, bigger), creating gargantuan datasets of billions of images that largely perpetuate the dominance of the same canonical—and mostly patriarchal—archetypes, is deeply biased and power-driven, almost blind to the challenges of equality and sustainability that the world is facing today (Dyer, 2016). As Joanna Zylinska points out, perhaps the biggest challenge regarding AI Ethics is not the technological development *per se*, but the social and economic structure that sustains it, "the

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

present-day capital-fuelled human thinking on AI itself [...] premised on a truncated, disembodied and yet so-very-gendered model of human subjectivity" (Zylinska, 2020, p. 41). Therefore, new approaches are essential, and we believe that visual motifs can be great tools to create and rethink AI and our relation with it, in aesthetically complex but also ethically critical ways.

2. Theoretical framework: visual motifs and AI ethics

To better understand the approach of our project, it is key to clarify what is a visual motif and how it connects with the broader fields of communication and information technologies. Close to their equivalents in music (André, 2007) and literature (Steiner, 1995), visual motifs can be defined as an iconographic pattern of cultural representation that is transmitted and reinterpreted through the history of images and that fosters narrative and emotional recognition. This premise means that pictorial and photographic representations shared by creators of various media —painters, photojournalists, filmmakers— integrate codes of creation and recognition of images based on specific formal variables. Their precedents can be traced back to the iconographic tradition (Panofsky, 1972), the study of visuality (Arnheim, 1969) and to the methods of cultural iconology (Warburg, 2010; Didi-Huberman, 2002), but more specifically, visual motifs have proven to be a very fruitful methodology to study cinema from a comparative perspective (Balló, 2000; Balló & Bergala, 2016). Capable of creating both "ways of representing" and "ways of seeing" (Berger, 2016), they are based on visual composition strategies such as the position of characters within the frame, certain gestures and uses of objects, movements and actions, the framing of shots, camera angle and point of view, or lighting and montage. In short, visual motifs encapsulate the search for a significant image, which contains a certain expressiveness in its formal disposition.

Theories of visual motifs are inspired by earlier works coming from transversal fields such as photography and critical theory (Sontag, 1977; Barthes, 1981; Berger, 2013), visual culture and iconology (Mitchell, 2005) or graphic design and social semiotics (Kress & van Leeuwen, 2006). In fact, members of our project have developed previous research on the connections between visual motifs and power structures, as in the case of political leaders (Salvadó et al., 2020) and economic imagery (Garin & Fernández, 2021), delving into gender and class divides. Previous literature comprises research about visual motifs and quantitative methodologies as well (Garin & Elduque, 2016), proving to what extent motif quantification can be useful to challenge hierarchical canons within art history, like rankings or best-of lists. Moreover, the malleability of motifs, their resurgence in very different works created by very different people, makes them useful tools to analyze photojournalism and audiovisual fiction comparatively, thus underlining the mixture of real and fictional layers that characterizes visual culture in the public sphere (Salvadó & Balló, 2023).

While leading researchers on AI Ethics still consider the field to be in an early stage, it is undeniable that "like all fields of knowledge, it is subject to power relations, political interests, and business objectives such as quarterly earnings goals" (Ebell et al., 2021). Evidence has been found in several contexts underlining that AI-based methods can strengthen pre-existing inequalities that severely affect certain groups of people (Fabbri et al., 2022). Consequently, there are key issues that are currently being discussed by the scientific community and that specifically relate to visual motifs. In particular, the comparative approach established by Daniel Chávez Heras (2019) between Film Studies and AI Studies is of great interest. Although Chávez Heras identifies new problems that arise as AI becomes integral to visual culture, his study goes beyond the common discussions about the ethics of AI art. The majority of these studies (Galanter, 2019; Stark & Crawford, 2019; Flick & Worrall, 2022) focus on ethics related

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

to generative AI, addressing issues such as copyright, artistic authorship, and negative practices of generated images such as deepfakes. Beyond the legislative domain, these studies do not propose anything new; rather, they revive an old debate that has been ongoing since the incorporation of mechanics into the arts, centered around the concept of "aura", introduced by Walter Benjamin in 1935, and expanded by key art historians like Didi-Huberman (2002). Studies like those by Joshua Krook (2024) bring this concept into the contemporary context of generative AI. Krook argues that there is a loss of the "aura" not only in the materiality of the artwork itself but also in the artist, who is no longer human. In this context, the responsible use of creative AI in these studies emphasizes the prioritization of human vision over computer vision. However, Chávez Heras' ethical perspective advocates for an interdisciplinary and horizontal approach to understand the relationships and complementations between AI and visual culture. His proposal involves, from a critical-technical analysis, the collaboration between art historians, film and media scholars, scientists, and engineers.

Following this collaborative work between humans and machines, numerous studies underscore the potential of AI as a tool for art analysis. Of particular significance are the contributions of Elgammal et al. (2018), who explored the challenges of classifying art styles using machine learning. They initially investigated whether the "eyes of the machine" could identify artistic styles canonized by art historians. The study yielded promising results: the learned representations consistently highlighted certain artists as the most distinctive representatives of their styles, quantitatively confirming art historian's observations. Similarly, David G. Stork (2023) contends that algorithms have increasingly become powerful tools in fine art studies. The algorithmic methods can offer art scholars, curators, and historians new avenues for interpreting art. Like Chávez Heras, Stork asserts that computer methods for art analysis must blend humanistic and scientific expertise to integrate technical analysis into culturally appropriate contexts for art studies. Hence, studies that detect patterns in large collections of artworks are also noteworthy (Gonthier et al., 2018; Shen et al., 2019; Madhu et al., 2020; Castellano & Vessio, 2021; Bernasconi, 2022). However, unlike these studies—whose primary objective is the recognition of figurative patterns in artworks belonging to specific artistic movements—our project, by applying the methodology of visual motifs to AI-based image recognition, finds greater affinities with other initiatives developed from an iconological perspective. In these cases, the goal is not to identify particular styles, but rather to reflect on the recognition of images through other images, across different artistic movements, media, and historical periods.

In this regard, it is relevant to highlight pioneering projects such as the Iconclass system, whose origins date back to the 1940s, when Henri van de Waal began developing a universal classification system for the thematic content of works of art. Today, Iconclass has become a valuable testing ground for AI and machine learning applications in the thematic analysis of images (Henri van de Waal Foundation, n.d.). Likewise, it is worth mentioning a recent project led by scholars Leonardo Impett and Franco Moretti (2017), who are developing a gesture recognition system for images based on Warburg's concept of *pathosformel*, with the objective, as they put it, of "moving from formulas to forms" (Impett & Moretti, 2017: 5).

3. Methodological approach: a curated and comparative dataset

A common agreement between AI Ethics experts is that, for the last decade, algorithms have been largely trained under a big data approach, meaning that quantity—millions and millions of images—clearly dominated over quality. Gargantuan amounts of data were used to train technology because they *could* be used, without many people asking if they *should*. And while this may sound a bit apocalyptic (to quote Umberto Eco's classic formula), the numbers are

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

there to prove it. Going back to the earlier example of Stable Diffusion, its datasets were scraped using automated image searches that privileged high-traffic websites (that is, mass-consumption) instead of visual archives based on expert sources. Thanks to Baio and Willison's analysis of a sample of 12 million images within the LAION-2B-EN set (a relatively small part of the total 2.3 billion images, but still indicative), we know that half of the images (47%) were sourced from only 100 domains, with the largest number coming from Pinterest, and the rest from user-generated content platforms such as Wordpress, Flickr, Smugmug, Blogspot, DevianArt or Tumblr (Baio, 2022). Therefore, given that the main criteria used to collect the dataset was "as many as possible" images from "as massive as possible" websites, it shouldn't come as a surprise (as we mentioned before) that the types of works, authors and characters represented are hyper-masculine, Anglo-Saxon, and topically recent. So, when choosing how to construct the image dataset for the *Visual Motifs Identification and Comparative Image Learning* project, we decided to base it on quality, not just quantity.

It is important to clarify, though, that the concept of a curated dataset is far from being new: prior to the development of machine learning, this was -and often still is- the method used to assemble physical archives in art museums, audiovisual libraries or cinematheques. Not by chance, some AI art projects, such as Mario Klingemann's work as artist-in-residence at Google Arts & Culture (Zylinska, 2020), were based on previously curated collections. Currently, there are innovative projects built around institutional archives, like collaborations between the University of Amsterdam and the Dutch EYE Filmmuseum, where models are trained with reels from their silent cinema collections, pushing the idea of AI creativity towards materiality (Celis Bueno et al., 2024). But in our case, instead of working with a pre-existing archive, we are creating, from scratch, a new dataset of images, coming from different art disciplines, that are purposely selected because they belong to -or fall within- specific visual motifs. Hence, our model is not only trained with data that's been curated by art historians, but what's more important, it is being trained with data *curated using a comparative methodology*, juxtaposing more traditional arts like painting, sculpture or architecture with modern arts such as cinema, photojournalism, comics or video games.

During the last two years, our research team has experimented with different sets of images organized around specific motifs, under a methodological approach that we call *Curated Comparative Dataset* (CCD), to feed our AI model with images selected from a qualitative perspective, instead of quantitative aggregates from the web. The selection of visual motifs is partially based on the iconographic work carried out within the Research Group CINEMA at Pompeu Fabra University, in previous funded projects like *Los motivos visuales en la esfera pública. Producción y circulación de imágenes del poder en España, 2011-2017* and *Fútbol y Cultura Visual en el Franquismo: Discursos de Clase, Género y Construcción Nacional en el Cine, la Prensa y los Noticiarios (1939-1975).* In the current stage, our dataset comprises up to 12,000 images that belong to 20 different visual motifs, including traditional ones such as *La Pietà* (the most well-known motif in art history) and others related to the public, political, or economic realms (like the *Leader Walks Alone*, a canonical visual depiction of leadership), as well as more transversal ones coming from the world of sport and popular culture (for instance, the *Signing Contract* motif). But, beyond the numbers, the most important aspect for us is how the fluid relations and limits between the different motifs foster a more complex, multi-layered and comparative approach to AI datasets.

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

Figure 1. Example of the *Shadow* visual motif with images from different mediums.

Source: Own elaboration.

In Figure 1, six images from six different mediums illustrate the common aesthetic matrix of the *Shadow* visual motif, exemplifying our comparative methodology: a classic vignette from *Lucky Luke* (comic book), a frame from the indie game *Limbo* (new media), a satirical portrait of Stalin (painting), a Saul Bass illustration (graphic design), a shot from Eisenstein's *Ivan Grozny* (cinema), and the opening credits of *Mad Men* (television). Therefore, by training our model with our *Curated Comparative Dataset*, we intend to challenge replicative and homogenizing approaches to art, because our goal is not to teach algorithms to imitate Saul Bass, or reproduce the chiaroscuro style of *Limbo* and certain noir films, but to develop technology capable of identifying and juxtaposing the visual motifs that traverse art history, in its many forms and periods. Our designed learning models are therefore trying to transcend the style and appearance of the image to understand the deeper meaning.

But how is such a complex dataset assembled, in practical terms? Due to the transversal composition of the research team, combining engineers and mathematicians with visual art historians, it is worth delving into the methodologies used to select, gather and compare the images during the first year. As far as the dataset is concerned, we have implemented two parallel methods: the discussion and critical selection of which visual motifs to work with and which images to select within each motif (1); and the actual collecting and tagging of specific images for each one of the twenty corresponding motifs (2).

3.1. *Visual motif selection and scope: core panels*

For decades, cultural historians have warned about the dangers of the unlimited growth of image consumption in late capitalist societies, from Susan Sontag's cautionary tales about the "market of images" and its insatiable logic (Sontag, 1977, p. 168), to Hito Steyerl's writings on the excess of visual data, the problem of quantities, with billions of images constantly competing to catch our attention and be shared and monetized (Steyerl, 2017). Joanna Zylinska has also underlined this problematic within the context of AI, when she talks about the "excess of productivity: an outpouring of seemingly different outcomes whose structure has been predicted by the algorithmic logic that underpins them" (Zylinska, 2020, p. 72). The notion that, currently, humans are exposed to too many images, too fast, may be debatable, but as far as our dataset is concerned, we wanted to work in a way that prevented the mere accumulation of images and fostered, instead, a critical and ethical debate about them. Following the comparative approach of Aby Warburg's *Atlas Mnemosyne*, we started every week in our

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

working calendar with a collective session where the center and the limits of each visual motif were discussed by the whole team, composing what we call *core panels*. These panels integrate approximately 40 images per motif by juxtaposing them on a 150cm x 100cm cork board, manually, without the use of digital or automated tools.

Figure 2. Example of a comparative *core panel* displaying the visual motif *Family Photo*.

Source: Own elaboration.

We believe that taking the time to do this with our own hands (instead of automating the process) is a key methodological choice, because printing, cutting out, and hanging images on the board to create the *core panel*, forces the team to share, discuss and rethink the very logics of each visual motif, and how it will affect the algorithmic models. Not only from an aesthetic standpoint, but also in ethical terms, since the team's discussions make us ponder the signifiers of class, race and gender evoked by the images, as well as the relations and tensions between them. In an age where the automated scraping of millions of images to feed AI seems to be the norm (Tiedrich, 2024), more humble and lo-fi methods may help balance the current obsession with big data and massive consumption. For instance, Figure 2 exemplifies the panel of one motif, *Family Photo*, with visual comparisons from very different artforms. But, methodology-wise, the key thing is that such collective debates (*which* images to print, *where* to place them in the board, *how* to connect them) enrich the technological process as well, given that the goal of these *core panels* is to push the limits and boundaries of all the visual motifs while favoring the AI model that we are simultaneously constructing and testing, to generalize and robustly consolidate the comparative learning. In this sense, it was essential in this phase to think about those approaches to the visual motif that, from an authorial perspective, offered a different dimension to what a priori could be a standard representation, typical of any image catalog.

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

3.2. *Image search, collection and classification*

While the aforementioned panels evoke interesting limits and possibilities from a qualitative standpoint, it is clear that for the proper development of learning-based technologies, bigger amounts of data are needed, and that is the second key process regarding the dataset. After creating the *core panel* of a visual motif, one per week, we proceeded to the search and collection of more images in order to train the algorithmic model: this was done by hiring a research technician with a visual arts background that, having participated in the collective discussions beforehand, could apply the aesthetic criteria established in them to the gathering of images. Thus, the choice of which images to include was not automated via data scraping (as in other AI datasets and models), but directly supervised by the researcher who selected, collected and tagged the data. As elaborate and time-consuming as that task may be, it has the advantages and the ethical safeguards of humanistic curation, with oversight throughout the whole process. This doesn't mean that, in future stages of the project, the dataset needs to remain as it is now, on the contrary, it could be expanded with the help of automated methods, but only after setting its foundations on a properly curated and comparative base. Currently, the dataset includes up to 1,000 images for each one of the twenty motifs, a number that varies depending on their aesthetic specificities.

When it comes to the classification of the different images, key for the weight they have in training the base model, we have applied five different categories: first, the few images belonging to the aforementioned *core panels*, that represent the maximum degree of aesthetic variability; second, a selection of images that we call *canonical* (around 10% per motif), more accurate in terms of the compositional and performative essence of the motif and its expressive sense; third, the largest group of *standard* images meant to increase the sample quantitatively; fourth, what we have called *multi-motif* images, which evoke more than one visual motif within the dataset; and fifth, a group of *red flag* images that, even if they belong to a motif (due to cultural or historical arguments), are less representative in terms of their visual layout, or may need the incorporation of additional priors or knowledge in the AI model. This functional classification is currently being complemented with a critical analysis of the dataset from an AI Ethics perspective, by tagging all the images to detect, quantify, and counter possible biases in terms of gender and race representation. This is something that is essential for the future development of the project, addressing equality and fair use.

4. Technological development: our baseline model

The compilation of a diverse visual motif dataset allows us to have a relatively large view of what each motif is and represents, across various art disciplines or styles, and using different visual conventions. Teaching an AI model this way of understanding images, however, is a complicated task. On one hand, the mathematical model and architecture has to be properly designed in order to achieve our goal. On the other hand, given the size of the current dataset, there is a possible risk of overfitting in the learning process: the model may not be able to generalize to images outside of the dataset. Typically, models that classify images are trained by repeatedly giving images from a large-scale labeled dataset as input, so that they can learn which visual elements define and distinguish each class (or in this case, motif), in the hopes that these aspects can then be found in new images, allowing for correct predictions. But since visual models usually need a lot more data to be able to automatically determine the exact elements of an image to focus on, and allow them to accurately distinguish between different classes, our first approach has been to use intermediate pre-existing tools that have distinct objectives, but which contain information that we can repurpose for our own goal.

Besides, and as a way to counter the risk of overfitting, the recent advances in self-supervised learning allow to train large neural network architectures with non-annotated data, thus enabling

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

the use of massive datasets. The so-called *foundation models*, e.g., (Brown et al., 2020; Devlin et al., 2019; Radford et al., 2021), are able to learn powerful semantic features by training on a broad dataset, and can then be adapted to solve certain tasks in a more reduced dataset. This is the main approach we have followed: firstly, giving images as input to the visual stream of a contrastive language-image pre-training (CLIP) model, (Radford et al., 2021; Ilharco et al., 2021; Sun et al., 2023), in order to extract features (encoded in a vector of 1024 values) that contain high-level semantic information; and secondly, giving these features as input to our own custom network (a multi layer perceptron, with two linear layers), that we train using an appropriate loss function with the comparative dataset we have previously curated, thus allowing the new model to select and interpret the appropriate features as relevant and representative of any given visual motif.

This process allows us to sift through the large amount of information describing the input image in CLIP's output, and focus instead on the key elements that distinguish the different visual motifs. Our model therefore learns to ignore features relating to style, artform, period or celebrities in the image, and anything else that CLIP may be conveying that is irrelevant, to comparatively identify the visual motif. Again, having a diverse and comparative dataset for training is essential, given that we are not working with the traditional semantic categories (like the aforementioned: styles, periods, names) but with more complex aesthetic and iconographic parameters, under a visual motif logic. Furthermore, the extra annotation performed on these images is used appropriately during the training phase. For instance, the previously called *red flag* images have a lower weight and therefore less of an impact on how our network parameters are modified when learning about the motif represented in that image. Similarly, what we call the *canonical* images (the most representative 10% within a motif) have a higher weight and the network is therefore more influenced by them. Last but not least, the *multi-motif* category of images in our dataset, that is, the ones representing several motifs, influence the network in a way that tries to improve the prediction output for all motifs that the image represents.

All in all, it is important to keep in mind that, being in the first year of a collaborative project with engineers, mathematicians and arts & humanities scholars, the technological tools and the ways to apply or modify them are still being discussed and tested among our team. For instance, during the year before obtaining the project funding, some object detection and identification algorithms (such as Detectron2 (Wu et al., 2019)) were used in order to experiment with different strategies for earlier and reduced versions of the dataset. Because, even if our visual motif approach is radically different in its conception of how objectual, figurative, chromatic, geometrical, situational and performative variables influence image creation and identification (see Section 3), previously existing strategies could be helpful to put our own model to the test, and explore new avenues.

5. Discussion and preliminary results

To properly contextualize and exemplify the project's results at the current stage, we will now discuss two main cases. The first one serves as a way to show how our model, which incorporates CLIP features that are further processed by our model to achieve comparative visual motif classification, clearly outperforms a classification directly obtained from CLIP features. To show it, we compare them on the same examples (Figures 3), while delving into the encountered classification and detection patterns. The second case will try to shed light on the artificial intelligence and aesthetic nuances of the visual-motif methodology, by discussing a selection of predictions and classification errors that exhibit how our model is learning and distinguishing between different visual motifs (Figures 4-5). We believe that both cases mobilize important discussions about not only the aesthetics but the ethics of images and artificial intelligence, that is,

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

to what extent humans and machines "take things for granted" when it comes to the limits, uses and horizons of art.

5.1. *Beyond the surface: iconographic logics vs. superficial appearances*

As is common practice in machine learning, we have divided our comparative dataset in three different sets: training, validation and test. The training set allows us to estimate the parameters of the model during the learning stage, using the validation set to ensure it is generalizing correctly. The samples in the test set are not seen at all during training and serve to assess the performance of the model and how it is able to generalize to new samples. We use 75% of the samples in the dataset for training, 10% for validation and 15% for testing. Each of the three subsets keeps the percentage of the five types of images (*core panel*, *canonical*, *standard*, *multi-motif* and *red flag*) present in the whole dataset, that is, the categories discussed earlier in the article (Section 3). Currently, the accuracy of the model's classification using the test set is 94.78%, meaning that 1,525 out of 1,609 images in the test set are automatically predicted as the correct visual motif. But beyond its efficiency, and the numbers, it is worth explaining *how* the model is challenging or complementing previous approaches to computer vision and image understanding.

Figure 3. Iconographic logics: our model is trained through the visual complexity of motifs such as *Signing Contract* (Left Half) and *Pep Talk* (Right Half) beyond the superficial appearances exemplified in nearby CLIP representations, which tends to group images according to obvious formal traits (Top Row).

Source: Own elaboration.

Although the features in CLIP provide powerful semantic information, the comparison between the two large images of Figure 3 (first row) shows to what extent its computed representation is not directly suitable for, and is completely different from, our visual-motif approach. Since CLIP's model relies on natural language description of images—not necessarily given by art historians—and based on our experimental analysis, we hypothesize that it is bound to "judge" images from their superficial appearances (e.g. image style, artform, type of objects in the image). In this particular case, and on a scale of -1 to 1 (with 1 meaning most similar), the similarity of their corresponding CLIP's representation (features) is 0.9131. The perceived resemblance can be explained by the office setting, the people in suits, the photogenic point light and the color palette; after all, the two images are stock photographs with models hired to pose as businessmen in a work environment. But what

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

about the *inner logics* and the ethical and aesthetic *differences* between them? Are the two large photos in Figure 3 only saying things to us (Barthes, 1972) about neutral business situations? With no more semantic depth? Or do they contain multi-layered, and conflictive meanings in more aesthetically-complex ways that could be identified by properly designed learning models?

When confronted with the same two photographs from Figure 3, that CLIP represents with very similar features, our visual-motif model predicts a radically different outcome: it reads the first one, with the three men seated around a piece of paper, as a clear example of the *Signing Contract* visual motif (our model predicts *Signing Contract* with 100% confidence); and it interprets the second one, with one character explaining things to other colleagues in a work meeting, as another motif, the *Pep Talk* or tactical briefing (also correct with 100% confidence). Therefore, the more superficial or at least aesthetically obvious criteria used to train other algorithms is challenged, because images are not related due to formal similarities, but differentiated and contrasted in relation to other visual motifs. In the first case, the apparently "innocent" image of businessmen signing a contract is compared (by our model) with the strong gender and power signifiers of a 1905 oil painting, George Sheridan Knowles' *Signing the marriage contract*, and a 2017 photojournalistic image of the US president Donald Trump signing a $110 million arms deal with Saudi Arabia, all of them displayed in the left half of Figure 3. From an iconographic perspective, the three of them contain similar elements that constitute the motif, such as characters signing and characters looking down to the paper, tables and chairs to seat, or the hand with pen gestures, but what's more interesting is *how* they can evoke very different meanings and issues, while juxtaposing diverse historical periods and artforms.

In the second case, what CLIP interprets as a neutral image of a presentation in a business environment, is analyzed by our model in a comparative fashion as the *Pep Talk* motif, relating its iconographic patterns with classic plan/briefing scenes from gangster and heist movies, in this case Quentin Tarantino's *Reservoir Dogs* (1992), and with an image coming from the world of sports, the basketball coach Pat Summit giving a halftime speech to her Tennessee Athletics players in the 2007 Final Four against North Carolina, both shown in the right half of Figure 3. So instead of taking images for granted, the goal here is that artificial intelligence becomes capable of comparing images that come from different environments and contexts critically (for instance, the *Pep Talk* has strong precedents in military imagery, that may be traced back to Napoleon's portraits), in order to relate complex meanings that go beyond the surface of images (in this case, the dark problematic relations between the worlds of business/money and mafia/gangsters, masterfully analyzed by the David Graeber in his indispensable study of debt (Graeber, 2011)). Of course, these are just six images from a dataset that contains thousands, and they are discussed here as preliminary results of a project that's still in its first year, but the main point is that, for our algorithm, the two large images in the top row of Figure 3 are not similar due to their superficial appearances, on the contrary, they are different because in spite of sharing some stylistic patterns (texture, color tone, format) they evoke different concepts, issues and histories.

Moreover, when it comes to the relation between ethics and artificial intelligence, we believe that a visual-motif approach to computer vision can foster a more critical analysis of images. In the *Signing Contract* examples, as well as in the *Pep Talk* ones, we can see that the perceived "neutrality" of visual communication is of course not neutral at all, given that all of them carry important gender imbalances. These patriarchal issues are well exemplified in the case of the marriage contracts and dowry practices portrayed by George Sheridan Knowles in his painting, where women were literally sold by signing a paper in front of a notary (notice the character's positions, gestures and gazes). Not to mention the fact that, in the *Pep Talk* example, it is very different to see a man or a woman in the position of command, controlling the situation via oral speech and body language, as in the case of

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

Pat Summit encouraging her team. But as well as unveiling gender signifiers, visual motifs can help to expose power mechanisms and inequalities (colonial or economic ones for instance), as seen in the traditional motif *Signing Contract*, that goes back to the blood pact between Faust and Mephistopheles in Goethe, or to the signature of the family trust in *Citizen Kane* (1941). Today, this motif is sadly exemplified by the many arm deals and weapon contracts still signed around the world, like the one between the US and Saudi Arabia in Figure 3: an image that suggests that what's being "sold to the devil" is not just a single soul, like Faust, but the rights and destinies of entire countries.

Last but not least, it is important to highlight that, while the CLIP model gives a very similar representation for the two principal images of Figure 3, 0.9131, because they belong to the same visual style and typology (stock photography), our model compares and interprets very different ones, as seen in the rest of Figure 3 with the juxtaposition of not only stock photography but a painting, a film, a sports broadcast, and photojournalism. So, in terms of the diversity and plurality of the machine learning process, the visual-motif approach not only processes varied signifiers but also a more diverse range of visual formats and artforms, in what we think is an ethical posture in itself: diverseness.

5.2. *Identifying initial errors to understand the AI model: context and gesture*

One of the most interesting aspects of our research, so far, is to study the reactions of the learned model in terms of which specific parts or fragments of images were more relevant to identifying each visual motif. We have determined which parts of the images the model focuses on when giving a particular prediction using heatmaps, that use a code to signal concrete areas of the image in hotter (red) or colder (blue) colors, the hotter the more meaningful. This is a common strategy that has been used in various deep learning applications (Selvaraju et al., 2017; Chefer et al. 2021), but what we find distinctive is to compare how the model interprets the visual motifs with how filmmakers, photographers and visual artists in general perceive them. That is, to contrast the more or less established conventions of cinematic *mise-en-scène* (Metz, 1974; Bordwell & Thompson, 1996; Martin, 2014) and of pictorial composition (Gombrich, 1950; Arnheim, 1982; Arasse, 1992) with the discriminating and prioritizing strategies of our model. While this is an ongoing research, and the results are preliminary, as a general rule we have found that the compositional choices and motif predictions of the algorithm are not at all less valuable than those of art experts or historians, on the contrary, it is precisely *thanks to* the perceived errors and mistakes of the technological tool that we are able to discuss and rethink the true nature—and the limits—of each visual motif. Because, as Franco Moretti said, in his defense of quantitative methods for the humanities: "problems without solution are exactly what we need in a field like ours, where we are used to asking only those questions for which we already have an answer" (2005, p. 26).

To better visualize this, we have selected another example as the article's last case study, the *Duel,* a canonical motif of tête-à-tête confrontations from Goya to film westerns or arcade video games. First of all, Figure 4 shows a selection of images from this motif, that haven't been seen by the model during training, along with their aforementioned heatmaps, corresponding to where a slightly modified version of our model focuses on in the image. This model uses a different version of CLIP as a backbone, yielding a lower accuracy (87.82%, 1413 correct predictions out of 1609), but allowing for the use of the heatmap tool. This is very helpful to see how the algorithm is reacting, and which zones or gestures within the image inform that process. Doing that, key questions arise: beyond the more traditional 19th-century version of dueling, what can be considered a *Duel* and what can't? Does it always have to include direct fighting, battle and some sort of weapon, or can it include sublimated or more civilized engagements like chess? Do only humans, or humans commanding animals, qualify as "having" a duel? Or can other practices such as bullfighting be

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

included? What about non-human forms such as *mecha* robots, superheroes or pixels in a computer game screen? Where is the frontier between the one-against-one duel and the group brawl or collective battle, and can the latter include the former? How can AI models discern between literal and metaphorical duels?

Figure 4. Some images representing the *Duel* motif in our dataset (left image in each case), the non-zero predictions of our model (bottom text), and the regions that correspond to a *Duel* (right image, using a heatmap code).

Source: Own elaboration.

Looking closer at the heatmap patterns given in Figure 4, we realize that the more symmetrical and dual the positions are, the easier it seems for the model to identify the visual motif, as seen in the 100% confidences in the *Duel* predictions for jousting images with horses (Figure 4e), boxing photographs (Figure 4h), as well as fighting video games (Figure 4g). However, when the iconographic patterns of confrontation take on more metaphorical overtones, the model does not always identify the motif, as in the examples of duels in a chess game where the loss of verticality and the seated positions are aspects that impair the identification of the motif

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

(Figure 4d and 4j). Something similar happens with video game duels (Figure 4b). Although the idea of confrontation between human being and machine is at the center of the imagery of this medium, its staging breaks with the traditional iconography and embraces a more abstract one due to the interference of technology in the rewriting of the duel motif. On that note, it is telling that one of the other examples with a lower accuracy (in this 10-image sample, since the dataset is much larger), corresponds to the image of a duel between a man and a bear (Figure 4f). Although in this case the verticality and the gesture of the man are typical of a body-to-body duel (with strong resonances with Francisco Goya's *Duelo a garrotazos*), the fact that the opponent is not another human being generates an identification of the motif by the model with a very low confidence (19%). Certainly, Figure 4 illustrates other gestural and positional variables, such as the difference between lateral and frontal compositions (with 96% confidence in *Duel* for the image with a runner and a horse (Figure 4i)). Other problematic images in this sample are those corresponding to duels in the form of an arm wrestling (Figure 4a and 4c), specifically two of them. In both cases, there is a misidentification of the principal motif, predicted as another of the visual motifs worked on in the research, *Handshake*. *A priori*, the error of the model can be understandable because of the gestures of the arms of the duelists in both images. However, when the duelists are a human being and a robot, the percentage of recognition is even lower, reinforcing the idea that the duel in its abstract dimension between the human being and the machine creates confusion, beyond the fact that the gestures are very similar in both images.

Figure 5. A selection of wrong predictions around the motifs *Duel, Brawl, Pietà, Handshake*.

Source: Own elaboration.

The last relevant aspect we want to discuss, sticking to the *Duel* example, is what happens when the model confuses a given visual motif with another one (Figure 5). The thumb war photograph, for instance, that had been labeled as a *Duel* within an earlier version of the

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

dataset, was interpreted as two other motifs, the *Handshake* (a canonical visual depiction of agreement, common in business and sport) and the *Hug* (a motif with a strong emotional charge), with accuracies around 40-60% depending on the prediction model. Such a "wrong" AI prediction, from the perspective of art history and visual culture, is extremely interesting because shaking hands or hugging are actions often performed before or after duels, and moreover, they tend to involve symmetric and dual compositions as well (two hands or two bodies as similar to two opponents), all of which points to the low accuracy of the arm wrestling image mentioned before. Secondly, Figure 5 confirmed the importance of differentiating between images with only two characters or those containing more of them (even if two of them are "dueling" in it), proven by the fact that the basketball image was read as another motif, the *Brawl*, with accuracies among 47 and 99, in a mistake that proved to be useful for the model.

Needless to say, these types of confusions are allowing us to refine the dataset and address other issues during the technological development. But what's more relevant here is that, even when the predictions of the AI were failing, or precisely because they *were* failing, we understood the importance of many gestural and contextual details like the direction of the characters' heads and gazes. That's what happened with the *Jason vs. Leatherface* comic book cover, that the model identified not only as a *Duel* or a *Brawl* but as a *Pietà* (with up to 54 accuracy). Because in the most canonical Pietàs, the two bodies portrayed tend to share a diagonal gaze, with the above character looking down to the suffering one beneath: and that is what the model read in that image, going beyond our expectations. So, as preliminary as they are, these results certainly question and enrich the notion of what a visual motif is or can be, putting to the test the visual capabilities of artificial *and* human intelligence.

6. Conclusions

In our humble attempt to offer new approaches to develop AI models, we believe that the use of visual motifs as a methodological and technical tool can contribute to improving the artistic diversity and the ethical conscience within the field. By assembling a comparative and curated dataset that juxtaposes multiple periods and artforms (from the *beaux arts* to films, comic books or video games), we create an AI model that goes beyond the identification of artists or styles. Instead, we focus on the iconological and meaning-making nuances that cut across visual culture as a whole, thus emphasizing the qualitative dimensions of images and their signifiers, as opposed to the quantitative dimension presupposed by current uses of big data. In addition to this, the qualitative dimension allows us to identify and avoid certain clichés of representation at the level of gender, race and class.

According to first results generated by the AI model developed, here exemplified with visual motifs such as *Pietà*, *Duel*, *Leader Walks Alone*, *Shadow, Family Photo, Signing Contract*, *Line* or *Pep Talk*, we find some examples that, although anecdotal at first sight, illustrate on one hand the ambiguity that certain human gestures can hide, and on the other hand, the challenges of generating technology capable of perceiving the expressive connotations of visual art. Even so, different tools such as heatmaps, model accuracies and prediction analyses are helping us to understand and improve the mechanisms used by our model to interpret a given image as one visual motif or another one (for instance, on which parts of the image it focuses its attention), in order to find a common visual language in which bringing an arm closer to shake hands or giving a hug, for example, are not interpreted primarily as threatening gestures, but as galvanic manifestations of affection and complicity.

Finally, the article defends that only through a real, open, and non hierarchical collaboration between engineers and art historians, can more balanced and responsible

ISSN 2386-7876 – © 2025 Communication & Society, 38(2) 218-235

approaches to AI and images emerge. The expression of an emotion (the so-called *pathos formel*) is at the heart of the concept of visual motif and at the basis of notions like "pseudomorphosis" (Panofsky, 1972) and "survival" (Didi-Huberman, 2002), that explain the perpetual circulation and transformation of images and the meanings and uses they acquire among us. With this in mind, we would like to contribute to a technological paradigm where the emotions and effects evoked by a single image are just as important as the thousands of visual data gathered to develop it, and where AI models reflect the diverse and critical backgrounds of the teams behind them.

References

André, E. (2007). *Esthétique du motif*. Paris: Puv Éditions.

Arasse, D. (1992). *Le détail*. Paris: Flammarion.

Arnheim, R. (1969). *Visual Thinking*. Los Angeles: University of California Press.

Arnheim, R. (1982). *The Power of the Center*. Berkeley: University of California Press.

Baio, A. (2022). Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator. *Waxy.org*.

Balló, J. (2000). *Imatges del Silenci.* Barcelona: Empúries.

Balló, J., & Bergala, A. (2016). (Eds.). *Motivos visuales del cine*. Barcelona: Galaxia Gutenberg.

Barthes, R. (1972). *Mythologies*. London: Paladin.

Barthes, R. (1981). *Camera Lucida.* New York: Hill and Wang.

Berger, J. (2013). *Understanding a Photograph*. London: Penguin.

Berger, J. (2016). *Modos de ver*. Barcelona: Gustavo Gili.

Bernasconi, V. (2022). *GAB - Gestures for Artworks Browsing*. En *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22 Companion)* (pp. 50–53). Association for Computing Machinery. https://doi.org/10.1145/3490100.3516470

Bordwell, D., & Thompson, K. (1996). *Film Art: An Introduction*. New York: McGraw Hill.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). *Language models are few-shot learners*. Advances in neural information processing systems, 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Castellano, G., & Vessio, G. (2021). Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Computing and Applications,* 33(19), 12263-12282. https://doi.org/10.1007/s00521-021-05893-z

Castells, M. (2009). *Comunicación y Poder*. Madrid: Alianza Editorial.

Celis Bueno, C., Chow, PS., & Popowicz, A. (2024). Not "what", but "where is creativity?": towards a relational-materialist approach to generative AI. *AI & Society* 39(2). https://doi.org/10.1007/s00146-024-01921-3

Chávez Heras, D. (2019) *Cinema and Machine Vision.* Edinburgh: Edinburgh University Press.

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 782-791).

Devlin, J., Chang, M–W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL-HLT. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Didi-Huberman, G. (2002). *La imagen superviviente.* Madrid: Abada.

Dyer, G. (2016). *1% Privilege in a Time of Global Inequality*. Berlin: Hatje Cantz.

Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J., MacIntyre, J., Madhamshettiwar, P., Maffeo, L., Matthews, J., Medsker, L., Smith, P., & Thais, S. (2021). Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics*, *1*, 131-138.

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M. & Mazzone, M. (2018). The Shape of Art History in the Eyes of the Machine. *Proceedings of the AAAI Conference on Artificial Intelligence,* 32(1). https://doi.org/10.1609/aaai.v32i1.11894

Fabbri, F., Croci, M. L., Bonchi, F., & Castillo, C. (2022). "Exposure inequality in people recommender systems: The long-term effects." Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, 194-204. https://doi.org/10.1609/icwsm.v16i1.19284

Flick, C., & Worrall, K. (2022). The Ethics of Creative AI. In: Vear, C., Poltronieri, F. (eds) *The Language of Creative AI. Springer Series on Cultural Computing.* Springer, Cham. https://doi.org/10.1007/978-3-031-10960-7_5

Foucault, M. (2009). *Vigilar y castigar*. México D.F.: Siglo XXI Editores.

Galanter, P. (2019). Artificial intelligence and problems in generative art theory. *Proceedings of EVA London 2019. BCS Learning & Development.* https://doi.org/10.14236/ewic/EVA2019.22

Garin, M., & Elduque, A. (2016). Quantitative meta-analysis of visual motifs throughout film history. *El profesional de la información*, 25(6), 883-892. https://doi.org/10.3145/epi.2016.nov.05

Garin, M., & Fernández, A. A. (2021). Imágenes y motivos visuales del poder económico español: la corte del Ibex y la crisis bancaria (2011- 2013). *Communication & Society*, 34(2), 333-350.

Graeber, D. (2011). *Debt: The First 1,000 Years*. Brooklyn: Melville House Publishing.

Gombrich, E. (1950). *The Story of Art*. London: Phaidon.

Gonthier, N., Gousseau, Y., Ladjal, S., & Bonfait, O. (2018). Weakly supervised object detection in artworks. *Proceedings of the European Conference on Computer Vision Workshops.* https://doi.org/10.48550/arXiv.1810.02569

Henri van de Waal Foundation. (n.d.). *Iconclass: An iconographic classification system*. Retrieved April 21, 2025, from https://iconclass.org/

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021). *Openclip*. Zenodo, https://doi.org/10.5281/zenodo.5143773.

Impett, L., & Moretti, F. (2017). *Totentanz: Operationalizing Aby Warburg's Pathosformeln* (Literary Lab Pamphlet No. 16). Stanford Literary Lab. https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf

Kress, G. R., & van Leeuwen, T. (2006). *Reading Images*. London: Routledge.

Krook, J. (2024) Art in the Age of Artificial Intelligence. http://dx.doi.org/10.2139/ssrn.4766175

Madhu, P., Marquart, T., Kosti, R., Bell, P., Maier, A. & Christlein, V. (2020). *Understanding Compositional Structures in Art Historical Images Using Pose and Gaze Priors: Towards Scene Understanding in Digital Art History*. European Conference on Computer Vision (pp. 109-125). https://doi.org/10.48550/arXiv.2009.03807

Manovich, L. (2020). *Cultural Analytics*. Cambridge, Massachusets: The MIT Press.

Martin, A. (2014). *Mise en Scène and Film Style.* London: Palgrave.

Metz, C. (1974). *Film Language*. Oxford: Oxford University Press.

Mitchell, W. J. T. (2005). *What Do Pictures Want?*. Chicago: The University of Chicago Press.

Moretti, F. (2005). *Graphs, Maps, Trees*. New York: Verso.

Morris, D. (2019). *Postures. Body Language in Art*. London: Thames & Hudson.

Panofsky, E. (1972). *Studies In Iconology*. New York: Routledge.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.* Proceedings of ICML. 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Salvadó, A., & Balló, J. (2023). (Eds.). *El poder en escena. Motivos visuales en la esfera pública*. Barcelona: Galaxia Gutenberg.

Salvadó-Romero, A., Fernández-Moreno, A. A., & Tedesco-Barlocco, B. (2020). "De Lincoln a Putin: el motivo visual del líder político caminando en los media españoles". En: Comunicación y diversidad. VII Congreso Internacional de la Asociación Española de Investigación de la Comunicación (AE-IC). Valencia, España, 28-30 de octubre, pp. 193-204. EPI SL. ISBN: 978 84 120239 5 4

Selvaraju, R. R., Cogswell, M., Das A., Vedantam, R., Parikh, D., & Batra D. (2017) *Grad-cam: Visual explanations from deep networks via gradient-based localization.* Proceedings of the IEEE International Conference on Computer Vision, 618–626. https://doi.org/10.48550/arXiv.1610.02391

Shen, X., Efros, A. A., & Aubry, M. (2019). *Discovering visual patterns in art collections with spatially-consistent feature learning.* Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 9278-9287. https://doi.org/10.48550/arXiv.1903.02678

Sontag, S. (1977). *On Photography*. London: Penguin.

Stark, L., & Crawford, K. (2019). The work of art in the age of artificial intelligence: What artists can teach us about the ethics of data practice. *Surveillance & Society*, 17(3/4), 442-455. https://doi.org/10.24908/ss.v17i3/4.10821

Steiner, G. (1995). *What is Comparative Literature?* Oxford: Clarendon Press.

Steyerl, H. (2017). *Duty Free Art. Art in the Age of Planetary Civil War*. New York: Verso

Stork, D. G. (2023). *Pixels & Paintings: Foundations of Computer-assisted Connoisseurship.* New York: John Wiley & Sons.

Sun, Q., Fang Y., Wu L., Wang X., & Cao Y. (2023). *EVA-CLIP: Improved Training Techniques for CLIP at Scale.* https://doi.org/10.48550/arXiv.2303.15389

Tiedrich, L. (2024). The AI data scraping challenge: How can we proceed responsibly? OECD. AI Policy Observatory.

Warburg, A. (2010). *Atlas Mnemosyne*. Madrid: Akal.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*, https://github.com/facebookresearch/detectron2

Zylinska, J. (2020). *AI Art.* London: Open Humanities Press.