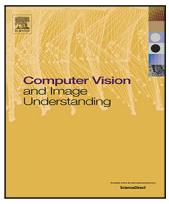
For a semiotic AI: Bridging computer vision and visual semiotics for computational observation of large scale facial image archives

Computer Vision and Image Understanding 249 (2024) 104187

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



For a semiotic AI: Bridging computer vision and visual semiotics for computational observation of large scale facial image archives

Lia Morra [a,*], Antonio Santangelo [b,1], Pietro Basci [a], Luca Piano [a], Fabio Garcea [a], Fabrizio Lamberti [a], Massimo Leone [b]

[a] Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Turin, Italy
[b] University of Turin, Via Verdi 8, 10124 Torino, Italy

ARTICLE INFO

Keywords:
Social media analysis
Image retrieval
Visual similarity
Semiotics
Digital humanities
Semantic image interpretation

ABSTRACT

Social networks are creating a digital world in which the cognitive, emotional, and pragmatic value of the imagery of human faces and bodies is arguably changing. However, researchers in the digital humanities are often ill-equipped to study these phenomena at scale. This work presents FRESCO (Face Representation in E-Societies through Computational Observation), a framework designed to explore the socio-cultural implications of images on social media platforms at scale. FRESCO deconstructs images into numerical and categorical variables using state-of-the-art computer vision techniques, aligning with the principles of visual semiotics. The framework analyzes images across three levels: the plastic level, encompassing fundamental visual features like lines and colors; the figurative level, representing specific entities or concepts; and the enunciation level, which focuses particularly on constructing the point of view of the spectator and observer. These levels are analyzed to discern deeper narrative layers within the imagery. Experimental validation confirms the reliability and utility of FRESCO, and we assess its consistency and precision across two public datasets. Subsequently, we introduce the FRESCO score, a metric derived from the framework's output that serves as a reliable measure of similarity in image content.

1. Introduction

In digital social networks, humans simultaneously produce and are exposed to an unprecedented amount of images. Many sociocultural practices are, as a consequence, changing the communicative power of digital representations and self-representations, most notably that of the human face. Digital image production has reached unprecedented levels in terms of quantity, pervasiveness, and potential for manipulation. The typical social media user spends more than two hours a day generating and scrolling through content, mostly in visual form Ortiz-Ospina (2019). Facebook, Instagram, Snapchat, Tinder, and other digital social networks are creating a digital world in which the cognitive, emotional, and pragmatic value of the imagery of human faces and bodies is arguably changing. However, researchers in the digital humanities are often ill-equipped to study these phenomena at scale. On the one hand, collecting and analyzing large amounts of images (so-called *visual big data*) require semiautomatic tools and techniques for visualization, exploration, and tagging (Manovich, 2020). Although the analysis of textual media has progressed extensively, the analysis of visual media is lagging behind. Existing platforms do not cater to the needs of digital humanities or focus on low-level visual features (Bocyte and van Kemenade, 2022). However, scholars in the digital humanities have developed sophisticated qualitative tools and techniques to interpret the multifaceted cultural significance of an image. There is a need to bridge these two approaches to reach insightful conclusions that are supported by adequate empirical evidence (Manovich, 2020; Bocyte and van Kemenade, 2022; Berlanga-Fernández and Reyes, 2024). Images on social media can be studied in many ways. In this article, we deal with the gaze we can cast on them, using the tools of visual semiotics (Polidoro, 2008; Eugeni, 2014; Pezzini and Spaziante, 2014; Mangano et al., 2018; Dondero, 2020; Corrain and Valenti, 2023). We believe that this discipline asks itself a series of very general questions, the solution to which is the basis of the way in which all other disciplines, from psychology to sociology, from anthropology to aesthetics, from philosophy to art history, relate to this type of content. Visual semiotics, in fact, questions how we assign meaning to them, knowing

* Corresponding author.
E-mail addresses: lia.morra@polito.it (L. Morra), antonio.santangelo@unito.it (A. Santangelo).

[1] All the authors collaborated in the theoretical discussion from which the draft of this article emerged. In particular, Antonio Santangelo wrote Sections 2.3 and 3, Lia Morra, Pietro Basci and Luca Piano the remaining sections.

https://doi.org/10.1016/j.cviu.2024.104187
Received 13 May 2024; Received in revised form 2 August 2024; Accepted 24 September 2024
Available online 26 September 2024

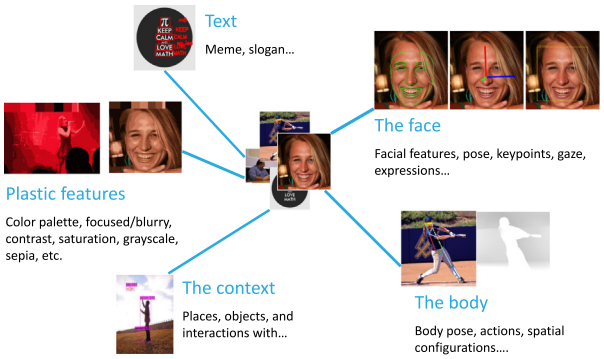


Fig. 1. The FRESCO (Face Representation in E-Societies through Computational Observation) pipeline extracts quantifiable traits from images using SOTA computer vision and deep learning tools. The traits are not limited to facial and body characteristics, but encompass interaction with the context and background, the presence of textual elements, and so forth. Such traits are categorized according to their plastic (color, forms), figurative (objects and actions) and enunciative (gazes and mutual placements) categories or traits, based on principles from structural visual semiotics.

full well that the interpretations we can produce are multiple. Nevertheless, it posits that any interpreter, when engaging with these forms of textuality, concentrates on certain specific fundamental components. These elements – of a plastic nature (shapes, colors, organization of space) or figurative (representations of the elements of the natural world), or related to the mechanisms which prompt the viewer to form a certain point of view on what is shown – are those that are usually considered pertinent by anyone who wants to assign a meaning to what they see in an image.

The core idea behind FRESCO (Face Representation in E-Societies through Computational Observation) was to develop a computational platform capable of bridging the gap between well-established semiotics principles and quantitative computational image interpretation techniques that could scale to hundreds and thousands of images. It builds on the tremendous advances in computer vision (CV) over the past decades and recognizes the potential of both established image processing techniques and the most recent foundational models in extracting *traits* from images, that is, characteristics that would be considered as potentially *pertinent* by visual semiotics scholars. As an example, and without loss of generality, such a platform could be used to cluster images produced by social media users based not only on their content, but also on their composition or their narrative structure. The FRESCO platform, by deconstructing images in a series of numerical and categorical variables, as depicted in Fig. 1, enables semioticians to take advantage of the extensive toolbox that the field of big data analytics and data mining has developed in the last decades to uncover novel and unexpected patterns from large visual collections.

In synthesis, our contributions are as follows:

- we introduce FRESCO, a computational framework that operationalizes structural visual semiotics in order to investigate the socio-cultural meaning of social media images at scale;
- we propose a practical implementation of the FRESCO framework and experimentally validate it on human-centered datasets to demonstrate the validity and usefulness of the proposed framework;
- we propose the FRESCO-score, a principled and transparent similarity measure based on the output of the FRESCO pipeline.

The remainder of the paper is organized as follows. Section 2 presents an overview of the related work. Section 3 provides some essential background on visual semiotics, while Section 4 illustrates the FRESCO computational pipeline in detail. Sections 6 and 7 present the experimental methodology and results, which are discussed in Section 8. Finally, Section 9 concludes the paper and suggests future work.

2. Related work

Many authors have investigated the interplay between computer vision and disciplines from the humanities, in particular between computer vision and art/media analysis (Datta et al., 2006; Hussain et al., 2017; Ye and Kovashka, 2018; Madhu et al., 2020; Wijntjes, 2021; Stork et al., 2021; Santos et al., 2021; Arnold et al., 2022; Yi et al., 2023), psychology (Strano, 2008; Ferwerda et al., 2015; Vilnai-Yavetz and Tifferet, 2015; Segalin et al., 2017b,a; Cucurull et al., 2018; Branz et al., 2020) and semiotics (Reyes and Sonesson, 2019; Ghidoli and Montanari, 2021; Pandiani and Presutti, 2023). In this section, the most relevant works to FRESCO and social media analysis in general are briefly reviewed.

2.1. Inferring personality from social media

Some studies show that it is possible, to some extent, to infer psychological traits from images published on social media, such as profile pictures (Branz et al., 2020; Ferwerda et al., 2015; Cucurull et al., 2018; Segalin et al., 2017b,a; Vilnai-Yavetz and Tifferet, 2015; Strano, 2008). For instance, Segalin et al. (2017b,a) investigated the ability of hand-crafted features and deep learning to infer self-assessed and attributed personality traits based on image features extracted from Facebook profile pictures. Their research suggests that images associated with a person can reveal some of their individual characteristics, such as their personality traits, with computerized assessment even outperforming human evaluation (Segalin et al., 2017b). In this type of study, social media users can be subjected to online questionnaires designed to self-assess personality traits, and then the classifiers are trained to predict the labels extracted from the questionnaire. The task is well defined with clear labels, and the problem is to extract/select relevant information. In FRESCO, we do not wish to make predictions on individual social media users, but we are rather interested in extracting multi-faceted, culturally relevant aspects of digital imagery.

2.2. Computational analysis in media and art history

Several computational platforms have been developed to analyze image archives in art history (Madhu et al., 2020; Yi et al., 2023; Wijntjes, 2021; Stork et al., 2021; Chen and Carneiro, 2015; Seguin et al., 2016; Elgammal et al., 2018), artistic/historical photography (Datta et al., 2006; Arnold et al., 2022; Männistö et al., 2022; Arnold and Tilton, 2020) and advertisement (Ye and Kovashka, 2018). Computerized tools can analyze, at scale and in a systematic fashion, large image archives. For instance, Elgammal et al. (2018) showed how machine learning can predict styles based on visual features and relate them to art history concepts. They show that representations learned by deep learning correlate with principles from art history and that predictions align with historical progression, thus providing a quantifiable verification of art historical theories. Other authors have focused on the use of machine learning to model and quantify image composition (Chen and Carneiro, 2015) or image aesthetics (Yi et al., 2023). Computerized analysis also allows art historians to establish links between different authors or artworks that may otherwise go undetected (Seguin et al., 2016).

While most of the above mentioned studies have focused on a few variables or on a specific analysis, in more recent years scholars have started to suggest that, in light of recent advances in computer vision and deep learning, a more extensive "visual grammar" could be operationalized and made accessible to the digital humanities scholar. In particular, Männistö et al. (2022) have proposed the AICE framework (Automatic Image Content Extraction) tailored to photography analysis. Their framework is based on the theoretical underpinnings of visual semiotics, and in particular the book "Image Grammar of Visual Design" by Kress and Leeuwen (1996), and Bell (2012)'s version Visual Content Analysis (VCA), a more practical and readily operable

adaptation of the original grammar which is particularly suitable for photographic analysis. In their book, Kress and Leeuwen (1996) presented an inventory of the major composition structures established as conventions in the history of visual semiotics and examined how they are used by contemporary image makers to generate meaning. Despite being developed independently and from different sources, FRESCO and AICE share many common characteristics. Both methods share the premises that visual semiotics provides a theoretical background to define a comprehensive lists of variables, which are mapped to state of the art computer vision and machine learning techniques. FRESCO is structured differently, grouping concepts according to different levels of analysis (plastic, figurative, and enunciational) initially defined by Greimas and refined by subsequent authors, as presented in greater detail in Section 3. We carefully reviewed the structure proposed in AICE to ensure that all the variables proposed therein are also covered in FRESCO. In addition, unlike Männistö et al. (2022) we provide a first practical implementation of FRESCO and go into greater detail into the accuracy and consistency of the extracted values, as well as practical issues that arise when trying to combine them into an overall similarity score.

2.3. Semiotics and computational analysis

Regarding the relationship between semiotics and computational analysis, the debate has first and foremost focused on how a discipline that originated at the intersection of philosophy and the social sciences, thus in the humanities, can dialogue with computer science and statistics. In this regard, a very important book is *Quantitative Semiotic Analysis* (Compagno, 2018), in which the issue of how to use tools for quantitative investigation is addressed when we set out to identify the meaning of a written or visual text, an activity that in the past has always been carried out using qualitative analysis methodologies.

Since signification is a deferential phenomenon, which people accomplish by linking signs to what those signs mean, thanks to codes that are not found in the texts themselves but in the minds of the interpreters and the culture they share, it has always proved more functional to assign the task of describing these mechanisms to a researcher and his or her ability to produce interpretations, as well as to imagine or recognize the interpretive logics of others. However, this inevitably reduces the extent of the content corpora that can be worked on, since this kind of investigation is constrained by methods of analysis that take a long time to be carried out. When one wishes to conduct studies on a very large corpus of texts, such as can be built in digital environments, it is necessary to make use of quantitative methods and tools. Confronting the various approaches to this problem in the various fields of the digital humanities (Moretti, 2005; Manovich, 2020), the authors of *Quantitative Semiotic Analysis* propose solutions that are in many ways similar to those we adopt in FRESCO: they emphasize, in fact, that computer systems must be designed to make use of semiotics to scrutinize their objects of analysis, recognizing their most significant elements and describing them in a way that is as functional as possible to enable the researchers who use them to best interpret the value of the data that these same systems produce.

The book edited by Compagno (2018), however, is also interesting for another reason: it deals, in fact, with a long series of theoretical problems raised by the encounter between semiotics and the techniques of quantitative investigation of large digital data corpora, but in its most applied part it deals only with the analysis of written texts. Only one article – that of Cholet (Cholet, 2018, ibid.: 101-121) – deals with the study of images, but with the technique of eye tracking. Thus, in practice, FRESCO's field of research is not considered, in this work. As is well known, after all, the computer tools that are used today to carry out quantitative semiotic analysis in the digital domain are mostly linguistic systems. So far, little has been done to reason about how to read and process images in an automated way using semiotics.

In this regard, the scarce available literature can be found first of all in the field of marketing. Ghidoli and Montanari (2021) reflect on some computer tools used to identify trends in consumer tastes. Using the Java SOM Toolbox framework, for example, O'Halloran (2015) analyzed large masses of images found online of young Japanese people having their pictures taken in their favorite clothes, producing a graph that can show how these can be divided into interrelated classes according to some logic that takes into account how fashion works in those latitudes (Owyong, 2009). Something similar, but at a broader level of generality, was done by the authors of ScenarioDNA.[2] In this case, different types of images found on social networks have been organized into concept maps, which allow them to be grouped into clusters of similar content, which derive their meaning because they differ from those found in other clusters that can be linked to them. By doing so, it is possible to conduct synchronic and diachronic analyses of the spread of these same contents. In addition, thanks to some network analysis tools, it is possible to understand how certain images spread in some networks of people rather than others. None of these systems focus, as FRESCO does, on face analysis, but the fact that they are beginning to be developed demonstrates the significance of our research project.

Another research field in which semiotics and computer science have often intersected is that of media and art history, as detailed in previous Section 2.2. Among those, the AICE framework for the analysis of photographic archives (Männistö et al., 2022), or the Distant Viewing tookit proposed by Arnold and Tilton (2019), are heavily based on semiotic principles. Our proposal pushes past these previous attempts by providing and validating an incomplete but extensive implementation of the proposed concepts.

Other works have relied on ideas that can be traced back to semiotics to design novel tasks for the computer vision community, focusing on the interpretation of "higher-level" semantic information or abstract concepts from still images (Pandiani and Presutti, 2023; Martinez Pandiani, 2024). For instance, Tores et al. (2024) proposed a computer vision task to detect the "male gaze" from video, that is, the objectification of women in video based on multiple cues such as camera placement and movement, gaze interactions, choice of clothing or nudity, posture, etc., many of which are also present in FRESCO. Other works have focused instead on quantifying image compositions in artworks based on pose and gaze information, focusing not only on subjects' pose, but also on composition lines established by the subjects' gazes (Madhu et al., 2020).

3. Background

As we have anticipated, FRESCO has been designed to allow scholars of visual semiotics to analyze the meaning of large amounts of images taken from the social profiles of people all over the world. Since the interpretation of this kind of content can differ depending on the research questions and the point of view of the researcher, our goal was to develop a computer system capable of reading the constituent elements of the images themselves which, according to the scientific literature, are usually taken into account to determine the meaning of the latter, whatever it is.

To identify these elements, we have used several texts, starting with Greimas' seminal essay entitled *Sémiotique figurative et sémiotique plastique* (Greimas, 1984), cited by many as the foundational work of modern visual semiotics studies (Corrain and Valenti, 2023). Then we turned to books on semiotic analysis of visual text in general (Polidoro, 2008; Eugeni, 2014; Dondero, 2020). Finally, we consulted works that deal with the semiotic study of photography (Mangano et al., 2018) and images on social media (Pezzini and Spaziante, 2014).

[2] https://www.scenariodna.com/

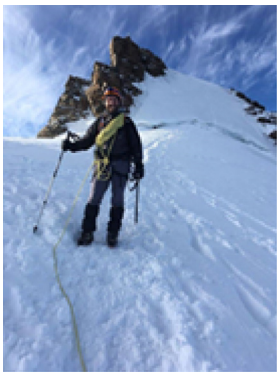


Fig. 2. The profile of a mountain climber.

Fig. 3. Picture of a man looking towards the clouds beneath him.

Fig. 4. A set of images with similar meaning.

All the authors of these articles and volumes agree that when we are faced with a figurative image such as those that, in most cases, are uploaded to our social profiles by people, one of the first interpretative actions is the recognition of the figures of the natural world that it reproduces: humans, animals, plants, objects, places, etc. It is also essential to recognize the actions of these subjects, which of them are active and which are passive, how they move, and what emotions they feel. All this serves to identify the main topic or topics of this image, but to do so and understand how the image frames the topic itself, it is also necessary to focus on the so-called "plastic" level. The latter comprises three categories of traits: eidetic, chromatic, and topological. The first category (eidetic) accounts for the shapes, lines, contours, dimensions, and symmetries of which the image is composed. The second (chromatic) for the colors, brightness, saturation, and textures. The last one (topological) for the spatial arrangement of all these contents, that is, what is above or below, right or left, in the center or in the periphery, in the foreground or in the background. All these elements, which compose the plastic structure of the image, contribute, together with the more figurative ones, to determine its meaning.

For example, as we have shown in a previous work (Santangelo and Morra, in press), in order to understand the meaning of Fig. 2, downloaded from the Facebook/Meta profile of one of the authors of this article, it is certainly important to understand that it depicts a man with mountaineering equipment, a mountain peak to climb, and a very steep slope made of snow and ice. But it is also essential to realize that the mountaineer covers only a small part of the image itself, which is otherwise occupied by the majesty of the natural environment; that he is more or less in the middle of the frame; that at the top of the image is the mountain top from which he has descended or on which he will soon climb, while a steep slope lies below; that the light illuminates his smiling face, giving it a warm hue in a context otherwise populated by cold colors. All these figurative and plastic elements help to communicate the happiness of being in the beauty of wild nature and being able to climb, feeling small but at the same time being the protagonist of a great adventure.

Another fundamental element, in order to understand the meaning of Fig. 2, as of any other image, is the construction of the observer's point of view, which Eugeni (2014) (op. cit.: 97-166) also calls *gaze system* or *watcher-looked system*. Eugeni himself argues that, depending on whether the latter is *basic*, *first-grade* or *second-grade*, it helps position the viewer of the image with respect to the latter and its contents, guiding how the image is "read" by the viewer. For example, speaking of the basic watcher-looked system, it is evident that a large painting of the face of Christ on the dome of a church is meant to be observed by much smaller people who are below, giving it a very specific meaning. However, a photograph on a social network page is composed to be observed from a very different position, which also generates greater engagement due to its communication style. On the other hand, speaking of the first-degree watcher-looked system, which again in photography refers substantially to the way in which the camera (or the camera of a smartphone) is placed, if we pay attention again to Fig. 2, the fact that it is taken from below and from a distance puts the observer in a position to appreciate the great steepness of the slope and the vertigo of the climb. Finally, coming to the *second-degree watcher-looked system*, which has to do with the direction of the gazes of the subjects represented, Fig. 3, also downloaded from the Facebook/Meta profile of one of the authors of this article, shows how important it is to look in the direction in which the protagonist of an image is looking, since there, evidently, lies a good part of the meaning of what the image itself wants to communicate.

A system like FRESCO must be able to recognize all the salient characteristics of a plastic, figurative nature and related to the construction of the gaze of the observer of an image, in order to combine them with those of the other images it processes. Such a system should aid the researcher expert in visual semiotics in identifying clusters of images, such as the three ones depicted in Fig. 4, that have many similar elements within them and, therefore, can be interpreted in the same way.

4. The FRESCO architecture

4.1. Conceptual design

The FRESCO architecture arises from a systematic mapping activity between concepts introduced in visual semiotics, introduced for the uninitiated reader in Section 3, and concepts and techniques developed in the context of CV. This mapping is informed by the authors' experience, by an extensive analysis of the current literature, as well as previous attempts from art history and photography history, such as AICE (Männistö et al., 2022). The result is presented in Table 1, in which the first two columns refer to the traits or categories commonly used in visual semiotics, while the last three columns denote their CV counterparts. Each trait was associated with one or more CV tasks that compute one or more quantitative measure: when this column is empty, it does not necessarily imply that the corresponding trait is not amenable to computerized analysis, but rather that to the best of the authors' knowledge the task has not been extensively tackled in the literature, and therefore suitable annotated datasets and models are not available. The last column illustrates the numeric output that is used to quantify the corresponding trait. In some cases, the output of a CV algorithm or model could be another image, such as a semantic segmentation map. To enable certain types of analysis, it would be preferable

to have more synthetic and numeric measures: for instance, if an image is reduced to a series of numeric or categorical measurements, it makes it easier to apply data analytics techniques to correlate them with other variables representing, e.g., socio-demographic measurements. For the traits currently implemented in FRESCO.v1, we therefore sought to define measurements that could be used for this purpose.

The first section of Table 1 represents technical information regarding the nature of the image: a photograph, an illustration, a map, a drawing, etc. The nature of this classification depends in part on the assumptions made about the archive under analysis. FRESCO was initially designed for the analysis of social media profile pictures, and thus to accommodate any type of imagery that a user may potentially select as a symbolic depiction of their face, while maintaining a focus on the face. Classifiers to distinguish different types of mediums can be trained with high accuracy (Cutzu et al., 2003; Wevers and Smits, 2020). For example, Wevers and Smits (2020) trained a CNN to distinguish historical photographs from several types of diagram. Alternatively, and without the need to define ad hoc categories, one can obtain a rough classification by employing a clustering technique on the features extracted from a pre-trained model. In the case of photographs, technical characteristics are often available from the file header, such as camera model, focal length, etc. (see Männistö et al. (2022) for a more thorough analysis of this aspect).

At the *plastic level*, the meaning of an image is constructed through a complex interplay of eidetic, chromatic, and topological categories. These plastic elements not only accentuate, but also sometimes contradict the figurative content, leading to nuanced interpretations and visual ambiguities. Many of the plastic categories identified in visual semiotics correspond to low-level image characteristics that have been studied in image processing and computer vision for decades.

Eidetic categories (1.1) pertain to the forms expressed in the images, through lines, contours, and textures. As discussed in Eugeni (2014), eidetic categories include first of all whether the image has mimetic or abstract qualities — that is, whether the image seeks to represent an existing objects or is rather an abstract image. Eidetic categories properties of the overall spatial composition such as the type of forms present (circular, square, etc.) (1.1.1), the symmetry of the composition and the main objects (1.1.2), the type of contours present (1.1.5), the main lines forming in the composition (1.1.6), etc. Many CV techniques have been developed to characterize the overall spatial composition of an image (Yao et al., 2012; Amirshahi et al., 2014; Wevers and Smits, 2020), such as evaluating the rule of thirds (Amirshahi et al., 2014). The presence of prominent compositional elements, such as diagonal line detection (1.1.7) and classification of spatial composition as vertical, horizontal or central (1.3.5), can be established borrowing from the field of computational photography (Yao et al., 2012).

In particular, spatial composition is of particular importance in the study of artistic photography (Yao et al., 2012), advertising (Wevers and Smits, 2020), and paintings (Dondero, 2020), in which the author of the image usually employs more sophisticated control over the composition of the image. In FRESCO.v1, we include only edge extraction among the existing tools.

Chromatic features encompass color (1.2.1), luminosity (1.2.2), saturation (1.2.3), and contrast, influencing emotional resonance and symbolic associations within the image. Chromatic features in FRESCO.v1 include global image features, such as palette and color histogram. Textural components, such as texture classification and clustering of image pixels based on textural and chromatic components (Bianconi et al., 2021) will be included in future work.

Topological features refer to spatial relationships, perspective and arrangement of elements, shaping the overall composition. Spatial relationships can be inferred from CV tasks such as object detection, semantic segmentation, panoptic segmentation, depth estimation and visual relationship detection. These tools produce as output spatial maps that can be directly used to, e.g., search for images with similar composition in terms of segmentation or depth map. However, as

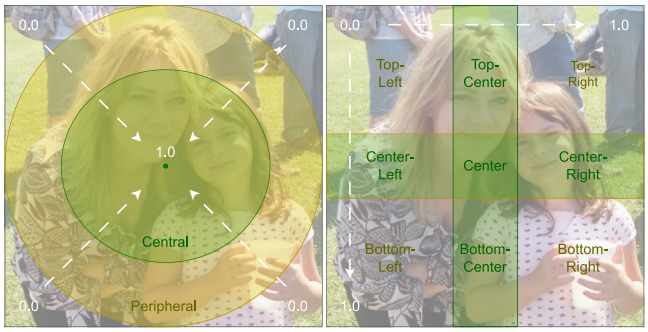


Fig. 5. We compute position of the centroid of each identified object or person with respect to the vertical and horizontal midlines, as well as the distance from the image center, to determine the position of each object or person with respect to the image frame. All positions are rescaled between 0 and 1 and thus are independent from image size.

stated before, we sought to define more concise and interpretable quantities to enable efficient indexing and comparison of large-scale image collections.

First, in visual semiotics the *spatial disposition* of each element can be determined with respect to the image frame, often represented in terms of oppositions (*central* vs. *peripheral* 1.3.1, *left* vs. *right* 1.3.2, *high* vs. *low* 1.3.3, *foreground* vs. *background* 1.3.4). In FRESCO.v1, we compute the position of the centroid of each identified object or person with respect to the vertical and horizontal midlines, as well as the distance from the center of the image (centrality), as illustrated in Fig. 5. Since positions are rescaled between 0 and 1, a value greater than or lower than 0.5 distinguishes between upper/left/peripheral and lower/right/central. As an approximation of whether an object is in the foreground or background, we compute the average depth by using a combination of panoptic segmentation and depth estimation. In visual semiotics, the figurative and plastic levels establish a complex interplay. At the computational level, this can be made evident by determining how we chose to partition the image into its constituent forms and elements. In FRESCO.v1, which focuses mostly on photography, elements are defined at the figurative level, through object detection and panotic segmentation. Other forms or elements could be extracted purely on the basis of plastic or compositional features (e.g., texture segmentation). This aspect should be kept in mind to account for future extensions.

Then, the spatial disposition of the elements with respect to each other is determined (1.3.5). In CV terms, these spatial relationships can be interpreted as a special case of the more general task of visual relationship detection (Cheng et al., 2022b). Spatial coverage (that is, the percentage covered by each class in a semantic segmentation) is also an indirect indicator of the spatial arrangement of objects (Männistö et al., 2022).

Plastic analysis also deals with how different forms and compositional elements interact with each other and how these interactions can shape the viewer's interpretation. Meaning is evoked by forms by assigning them qualities, which derive both from their internal characteristics and, above all, from the network of *spatial, temporal and cooperative or contrastive relationships* with other forms and the surrounding space (Eugeni, 2014). These connections can be established among *adjacent* (1.4) or *distant* (1.5) forms based on their *similarities* (e.g., same shape) or *differences* (e.g., dark vs. light). These connections are independent of the figurative content of the respective forms and may thus reinforce or redefine the interpretation that may be formed based on the figurative content alone.

Finally, all plastic elements contribute to the overall configuration (1.6), which can be balanced (predominantly static) or unbalanced (predominantly dynamic) (1.6.1).

The *figurative* level is concerned with the main topic (2.1), persons, objects, scene and setting (2.2), movement (2.3), actions (2.4) and emotions (2.5).

Table 1
Face Representation in E-Societies through Computational Observation (FRESCO) Computational Framework. The table maps semiotic-inspired category or variable with the corresponding computer vision task, if available, and the values that the variable can assume (either output by the computer vision task, or appropriately summarized). A ✓indicates variables that are included in the current FRESCOv1 implementation and the experimental validation in this paper.

	Semiotic category	Computer vision task		Values
0	**Technical**			
0.1	Image type	classification/clustering	✓	Photograph/illustration/map/..
1	**Plastic level**			
1.1	**Eidetic categories**			
1.1.1	Form			
1.1.2	Simmetry			
1.1.3	Mimetic/abstract			
1.1.4	Geometric/non-geometric			
1.1.5	Kind of contour			
1.1.6	Lines	Edge extraction	✓	Line map
		Diagonal element detection		diagonal_ulbr, diagonal_urlb, …
1.2	**Chromatic categories**			
1.2.1	Color	Palette estimation	✓	Palette
		Color	✓	Grayscale/color
		Color distribution	✓	Histogram
1.2.2	Luminosity	Brightness estimation	✓	Brightness
1.2.3	Saturation	Saturation estimation	✓	Saturation
1.2.4	Texture	Texture classification		
1.3	**Topological categories**			
1.3.1	High/low	Object detection	✓	Position of each object centroid w.r.t. the image midline
1.3.2	Left/right	object detection	✓	Position of each centroid w.r.t. the image midline
1.3.3	Central/peripheral	object detection	✓	Centrality ratio of each object
1.3.4	Foreground/background	panoptic segm. + depth	✓	Avg. depth value of each person
		panoptic segm. + depth	✓	Avg. depth value of each object
1.3.5	Spatial disposition of forms	visual relationship detection		scene graph ('X' left of 'Y', etc.)
		semantic segmentation	✓	semantic segmentation map
		semantic segmentation	✓	spatial coverage (percentage covered by each class)
		spatial composition class		vertical, horizontal, centered
1.3.7	Dynamization of forms			
1.4	Links between adjacent forms			
1.4.1	By similarity			
1.4.2	By confrontation			
1.5	Links between distant forms			
1.5.1	By similarity			
1.5.2	By confrontation			
1.6	Overall configuration			
1.6.1	Static vs. Dynamic	Classification		
2	**Figurative level**			
2.1	**General**			
2.1.1	Main topic	Classification/Clustering	✓	Person/animal/object/environment/event/...
		Image tagging	✓	tags
2.1.2	Salience	Salience estimation		salience map
2.2	**Persons/objects/scene**			
2.2.1	**Characteristics of people groups**			
2.2.1.1	Number of people	Face/person detection	✓	0/1(single)/2(couple)/3-6(small group)/7-12(medium group)/13-30(large group)/31-(crowd)
2.2.1.2	Number of groups	Gaze estimation/Social distance estimation		1/2/3+
2.2.1.3	Group typology	Gaze estimation/Social distance estimation		Unfocused/ common focused/jointly focused/
2.2.1.4	Group type	Classification		Family/friends/sport team/...
2.2.1.5	Atmosphere	Classification		Casual/formal/intimate/festive/...
2.2.2	**Characteristics of each person**			
2.2.2.1	Status	Main character recognition		Main Character (MC)/Side Character (SC)

(*continued on next page*)

Table 1 (*continued*).

	Semiotic category	Computer vision task		Values
2.2.2.2	Age	Age estimation	✓	Baby/child/young/adult/old
2.2.2.3	Gender	Attribute prediction	✓	Male/female/other
2.2.2.4	Identity	Face recognition		Name
2.2.2.5	Ethnicity	Attribute prediction	✓	European/Asian/African/...
2.2.2.6	Height	Height estimation		Short/average/tall
2.2.2.7	Weight	Weight estimation		Thin/average/fat
2.2.2.8	Occupation	Classification		Doctor/police/cook/pilot/
2.2.2.9	Role	Visual relationship detection		Child/mother/friend/neighbor/...
2.2.2.10	Nudity	Nudity detection		Nude/partially nude/clothed
2.2.2.11	Physical condition	classification		Healthy/sick/wounded/dead/...
2.2.2.12	Clothes	Attribute classification/Object detection	✓	
2.2.2.13	Clothing style	Style clustering		
2.2.2.14	Face and head accessories	Attribute prediction	✓	Glasses/jewellery/hat/..
2.2.2.15	Facial attributes	Attribute prediction	✓	Eyebrows/nose type/double chin/cheeks ...
2.2.2.16	Facial expressions	Attribute prediction Facial keypoints	✓	Smile/frown/...
2.2.2.17	Hair attributes		✓	Beard /Hair length/Hairline/Bangs/Sideburns...
2.2.3	**Objects**			
2.2.3.1	Status			Main motif (MM)/side motif (SM)
2.2.3.2	Category	Object recognition	✓	Animal/object
2.2.3.3	Text in image	Text recognition/OCR	✓	Text in image
2.2.4	**Settings/events**			
2.2.4.1	Scene class	Scene classification/Tagging	✓	Urban/rural/forest/hospital/school/
2.2.4.2	Privacy			Private/semi-public/public
2.2.4.3	Indoor/outdoor	scene classification	✓	Indoor/outdoor
2.2.4.4	Man-made/natural	scene classification	✓	Man-made/natural
2.2.4.5	Event	event recognition		
2.2.4.6	Location	location recognition/landmark detection		Location
2.2.4.7	Time of day	classification		Morning/day/evening/night
2.2.4.8	Time of year	classification		Winter/spring/summer/autumn
2.2.4.9	Weather	classification		Sunny/cloudy/raining/snowing/...
2.3	**Movement**			
2.3.1	Type of movement			Blocked/contracted/articulated
2.3.2	Visibility			Hidden/manifest
2.4	**Action**			
2.4.1	Single action	action recognition		classification
		body pose	✓	pose
		caption generation	✓	textual description
2.4.2	Aggregate of actions	visual relationship detection		scene graph
2.4.3	Narrative			
2.5	**Emotions**			
2.5.1	Intensity	Arousal regression	✓	arousal
2.5.2	Emotion recognition	emotion classification	✓	happy/neutral/fear/sadness/disgust
2.5.3	Emotional valence	valence regression	✓	valence
3	**Enunciational level**			
3.1	**Basic watcher-looked system**: the viewer			
3.1.1	position of the viewer	panoptic segm. + depth	✓	distance of the main subject(s) from the camera
3.1.2	position of the viewer	panoptic segm. + depth	✓	distance of the main character(s) from the camera
3.1.3	position of the viewer	horizon line estimation		position of the horizon line (frontal/from above/from below)
		scene classification	✓	indoor/outdoor
		framing	✓	portrait vs. scene
3.1.4	position of the camera	camera pose estimation		
3.2	**First-grade secondary watcher-looked system**: the observer subject			
3.2.1	position of the observer	head pose	✓	angle (yaw/pitch/roll)

(*continued on next page*)

Table 1 (*continued*).

	Semiotic category	Computer vision task		Values
3.2.2	position of the observer	body pose	✓	shoulder/hip angle (frontal/rotated left/rotated right)
3.2.3	position of observer	gaze direction	✓	angle (yaw/pitch)
3.2.4	position of the observer	presence/absence of perspective		classification
3.2.5	position of the observer	vanishing point regression		vanishing point positions wrt the image frame
3.3	**Second-grade secondary watcher-looked system**: indicators/bystanders; insignias and epigraphs			
3.3.1	bystanders	main character detection + gaze detection		
3.3.2	indicators	main character detection + action recognition + pose estimation		
3.3.3	insignias	object detection		
3.3.4	epigraphs			
3.4	**Spatial relationship**			
3.4.1	of secondary watcher-looked systems (first and second grade)			
3.4.1	of secondary watcher-looked systems (first and second grade) vs. basic watcher-looked system			coincident/rotated left/rotated right/opposite
3.4.2	of first grade secondary watcher-looked system vs. second grade secondary watcher-looked system			

The characterization of individual and groups of people is partially based on AICE (Männistö et al., 2022), which in turn is based on the work of Kress and Leeuwen (1996). We do not distinguish explicitly between attributes of the main character and of the side character, but assume that the categorization is available for each character, and expand the characterization to include attributes available in pre-trained facial attribute extractors (Zheng et al., 2022).

Eugeni (2014) distinguishes among movement (2.3) and actions (2.4). The movement categories pertain to how the image, which by nature is static, captures the evolving temporal dynamics of the scene. Over time, different strategies have been evolved to suggest how the scene depicted articulates in time, so that the view can evoke the temporal continuity that the still image cannot physically represent. Such techniques can be differentiated based on whether the image represents one or more instants in time within the same frame, which are classified by Eugeni (2014) in *blocked, contracted, or articulated* (2.3.1). The resulting configuration may shun realism in favor of making the articulation of movement *manifest* in the image (2.3.2); otherwise, the articulation of movement is assumed to be *hidden* in the presentation (2.3.2). While semiotics deals with all forms of still images, not only photographs but also paintings and illustrations, in FRESCO.v1 we concentrate on photographs and especially social image images, which are likely going to represent a single instant in time (blocked). Actions, on the other hand, refer to the semantic interpretation of the depicted gestures and interactions (2.4). *Single actions* (2.4.1) can be associated with CV tasks such as pose estimation and action recognition. In the case of *aggregate of actions* (2.4.2) or narratives (2.4.3), estimating the scene graph or detecting visual relationships would be necessary to differentiate the gestures performed by different characters and capture interactions among them. Finally, at the figurative level, we measure the *intensity* (2.5.1) and *class* of the *emotions* (2.5.2) expressed by the characters depicted in the image. Some works in the CV literature have also investigated how to determine emotional valence (2.5.3), that is, the emotion aroused by the image (Lu et al., 2016) in the viewer.

To conclude the figurative level, we also included as part of FRESCO.v1 image tagging (Huang et al., 2023) and visual captions (Hu et al., 2022). These models have the advantage of being trained on extremely large-scale datasets and thus were designed to achieve strong open-set capabilities, which are essential in the context of social media. At the same time, textual descriptions cannot be easily mapped to a specific figurative element, and thus may pose some issues when interpreting the results.

At the *enunciational level*, we focus in particular on the construction of the point of view of the *spectator* (*basic watcher-looked system* 3.1) and the *observer* (*first grade secondary watcher-looked system* 3.2). The former, in photography, essentially coincides with the position of the camera. It can be reconstructed on the basis of aspects such as perspective (horizon line and vanishing points), the position of the camera with respect to the scene, and the distance between the main subject(s) and the viewer. It is important to distinguish close-up and portraits from indoor and outdoor scenes, since the position of the spectator cannot always be clearly defined and is inferred based on different compositional cues depending on the type of image.

The way a photograph is framed, and therefore what is not shown as well as what it is shown, is of paramount importance to shape its interpretation. By *observer*, we denote a character that is explicitly never depicted, but implicitly assumed by a composition. The position of the observer can be inferred by the *pose* of the characters (*body pose* 3.2.2 and *head pose* 3.2.1), and most importantly by the *direction of the gaze* (3.2.3). The *relative position* of the spectator and observer (3.4.1) will elicit involvement or detachment in the viewer, depending on whether they coincide or differ. The presence of bystanders, indicators, insignias, and epigraphs [3] further helps to guide the viewer in the correct interpretation, with a higher level of guidance reflecting in a greater sense of participation, especially in artistic composition (Eugeni, 2014).

Several variables evaluated in FRESCO involve estimating the distances between the observer (the camera, in the case of a photographic

[3] In semiotics, bystanders and indicators refer to secondary characters that are looking, pointing or otherwise directing the viewers' attention to the main focus of the scene. Insignias and epigraphs are objects (such as mirrors) or spatial elements of the scene (such as the presence of doors, or the direction of the light) with similar function.

image) and the subject(s) depicted, as well as among the subjects depicted in the images in the case of multiple subjects. We estimate the distance of the main character combining the depth map and the panoptic segmentation considering only the "person" category. Similarly, we compute also the distance of main subject taking into account all the "things" categories, which includes only countable objects, since the primary theme in a photo may not necessarily be a person. Interpersonal distances are shaped by our sensory-motor possibilities (e.g., whether we can touch, hear, or smell another person) but are also influenced by social and cultural conventions; hence, they carry with them a plethora of implicit messages. Moving from the seminal works of Hall (1966), one of the cardinal findings of proxemics dictates that people tend to organize the space around them in terms of four concentric zones (intimate zone, casual personal zone, social zone, and public zone) associated with increasing degrees of intimacy and interactions. This classification forms the basis for subsequent works in visual semiotics (Bell, 2012), as well as in computational visual proxemics or Visual Social Distancing (VSD) estimation, that is, approaches that rely on cameras and other imaging sensors to analyze the proxemic behavior of people (Cristani et al., 2011, 2020).

4.2. Implementation

FRESCO relies on a collection of open-source, off-the-shelf CV models representing the state of the art in their respective tasks. Although we recognize that potentially more accurate results could be achieved using cloud-based commercial APIs, for the sake of privacy, reproducibility, and transparency, open implementation was preferred (Santangelo and Morra, in press).

Built-in models. FRESCO.v1 includes the following models, whose key details are summarized in Table 2. Face detection (1) is obtained using RetinaFace (Deng et al., 2020) with a ResNet50 backbone. The face mesh (2) is obtained from MediaPipe (Lugaresi et al., 2019), while the body pose (3) is obtained using PifPaf (Kreiss et al., 2019). Head pose (4) is estimated from 6DRepNet (Hempel et al., 2022), while gaze direction (5) is extracted using 3DGazeNet (Ververas et al., 2022) using the InsightFace implementation. Continuous levels of valence/arousal (6) and emotion category (6) are estimated using EmoNet (Toisoul et al., 2021), while 40 facial attributes (7), corresponding to those available in the CelebA dataset (Liu et al., 2015), are extracted using FACER (Zheng et al., 2022). Age (8), gender (9), and ethnicity (10) are estimated using the DeepFace (Serengil and Ozpinar, 2021) framework. Depth estimation (11), edge detection (12), object detection (13), OCR (14), semantic segmentation (15), panoptic segmentation (16) and caption generation (17) are derived through PRISMER (Liu et al., 2023) and its associated expert models (Ranftl et al., 2021; Poma et al., 2020; Zhou et al., 2022; Liu et al., 2018; Cheng et al., 2022a). Image tags (18) are extracted using the Recognize Anything Model - RAM++ (Huang et al., 2023). Scene classification (19) is performed using a VGG model trained on Places365 (Zhou et al., 2017). Chromatic information (20) is extracted using established image processing techniques, while simple geometric measures are implemented custom. An example of the output of these models can be seen in Fig. 6.

Structured data extraction. We tightly combined the output of these pre-trained models with geometric properties and/or image processing methods to extract the information described in Section 4.1 and which correspond to the items marked by ✓in Table 1.

At the *Plastic level*, the *Eidetic features*, which in the current version includes only the line map (1.1.6), are obtained from the edge detector (12). *Chromatic features* (1.2.1-3) are extracted using image processing techniques (20). For the *Topological features*, those related to spatial disposition (1.3.1-3) are obtained from a direct comparison of the centroids' positions, derived from the bounding boxes found by the object detector (13), and the image area, resulting in three different positional ratios ranging in [0,1]. These values can be discretized to obtain the position of each object as Top/Center/Bottom, Left/Center/Right and Central/Periperal as exemplified in Fig. 5. In the current implementation, each centroid position is discretized considering the image area divided in three bands in the proportion 40:20:40, both vertically and horizontally, for the first two measures. An object is instead considered central if its centroid falls within an ellipse having semiaxes equal to 60% of the image semiaxes. All these thresholds were chosen empirically and can be adjusted through parameters. The depth positions (1.3.4) are obtained by combining the output of the panoptic model (16) and the depth estimator (11). Specifically, the depth maps of each person and object detected in the image are isolated by masking the whole depth map with each instance segmentation map found by the panoptic and then averaged to obtain the average depth for each instance. The background average depth is obtained by averaging the remaining part of the depth map once all pixels corresponding to objects and people are removed. For spatial disposition (1.3.5), we estimate spatial coverage by computing the percentage of pixels covered by each class on the semantic segmentation map (15).

At the *Figurative level*, the main topic (2.1.1) is estimated using the image tagging model (18). The number of people (2.2.1.1) is derived from the output of the object detector (13). Even though the face detector (1) is highly accurate in its task, we decided to rely on the object detector to also consider people with occluded faces or photographed from behind. The characteristics of each person (2.2.2) are obtained by running on each face crop extracted by the face detector (1) the respective model for each task including the age (8), gender (9), ethnicity (10), face attributes (6). The object category and the text in the image (2.2.3.2-3) are obtained using the labels found by the object detector (13) and the OCR model (14). Scene characteristics (2.2.4.1 and 2.2.4.3-4) are inferred from the output of the scene classification model (19). For action (2.4.1), the body pose and the image caption are extracted using (3) and (17). Similarly to the single person characteristics, the emotions (2.5.1-3) measures are evaluated on each face crop extracted by the face detector (1) and using the emotion estimator (6) that returns both continuous values (valence/arousal) and emotion category.

At the *Enunciational level*, for the *Basic watcher-looked system*, the distances from the camera (3.1.1-2) are estimated following an approach similar to the one adopted for (1.3.4). The framing (3.1.3) is instead obtained by computing the ratio between the area of the largest face crop found by the face detector (1) and the entire area of the image. In the current implementation, we consider an image as a portrait if the face crop covers more than 30% of the total image or as a scene otherwise; the threshold can also be adjusted through a parameter. For the *First-grade secondary watcher-looked system*, all angles (3.2.1, 3.2.3) related to head pose (yaw, pitch, roll) and gaze (yaw, pitch) are estimated on each face crop extracted by the face detector (1), by running the models for the head pose estimation (4) and gaze direction (5).

As a result, we obtain two sets of image-level functions $\mathcal{F}_t := \{f_t^k(\cdot)\}$, with $k = 1...K$, and object-level functions $\mathcal{G}_t := \{g_t^l(\cdot)\}$, with $l = 1...L$, and $t \in \{\texttt{plastic}, \texttt{figurative}, \texttt{enunciational}\}$, that take as input the whole image or at each single object detected in the image, respectively, for each level of analysis t, and that can be exploited to compare the content of the image as discussed in the following section. We make the implementation available at https://gitlab.com/grains2/fresco.

5. The FRESCO similarity score

In this Section, we define FRESCO-Score, a similarity measure that leverages FRESCO, and specifically all the measures available in FRESCO.v1 as highlighted in Table 1. It represents an estimation of how closely two images represent the same content at the plastic, figurative, and enunciational levels. Unlike feature-based similarity metrics (Ramtoula et al., 2023; Hessel et al., 2021; Heusel et al., 2017),

Table 2
Models included in FRESCO v1 implementation.

Task	Model	Dataset (train)	Dataset (test)	Performance (expected)
Face Detection	RetinaFace (ResNet50) (Deng et al., 2020)	WIDERFACE (train)	WIDERFACE (val)	mAP: 96.5%(easy), 95.6% (medium), 90.4% (hard)
Face Mesh	MediaPipe (Lugaresi et al., 2019)	Private	Private	IOD MAD: 3.96%
Head Pose	6DRepNet (Hempel et al., 2022)	300W-LP	AFLW2000	Yaw: 3.63, Pitch: 4.91, Roll: 3.37, MAE: 3.97
Gaze Direction	3DGazeNet (Ververas et al., 2022)	Gaze360 (train)	Gaze360 (test)	Gaze error (degrees): 9.6
Emotion Estimation	EmoNet (Toisoul et al., 2021)	AffectNet (train)	AffectNet (test)	Expression Acc: 0.75 Valence CCC: 0.82, PCC: 0.82, RMSE: 0.29, SAGR: 0.84 Arousal CCC: 0.75, PCC: 0.75, RMSE: 0.27, SAGR: 0.80
Face Attribute Estimation	FACER (Zheng et al., 2022)	LAION-Face-20M + CelebA (train)	CelebA (test)	Acc: 92.1%
Age Estimation	DeepFace (Serengil and Ozpinar, 2021)	IMDB-WIKI	IMDB-WIKI	MAE: 4.65
Gender Estimation	DeepFace (Serengil and Ozpinar, 2021)	IMDB-WIKI	IMDB-WIKI	Acc: 97.44%, Precision: 96.29%, Recall: 95.05%
Ethnicity Estimation	DeepFace (Serengil and Ozpinar, 2021)	FairFace (train)	FairFace (test)	Acc: 68.0%
Image Tags	RAM++ (Swin-L) (Huang et al., 2023)	COCO + VG + SBU captions + Conceptual Captions + Conceptual 12M	OpenImages, ImageNet-Multi, HICO	Tag-Common mAP: 86.6 (OpenImages), 72.4 (ImageNet-Multi) Tag-Uncommon mAP: 75.4 (OpenImages), 55.0 (ImageNet-Multi) Phrase-HOI mAP: 37.7 (HICO)
Scene Classification	VGG-Places365 (Zhou et al., 2017)	Places365 (train)	Places365 (test)	Top-1 acc: 55.19%, Top-5 acc: 85.01%
Body Pose	PifPaf (Kreiss et al., 2019)	COCO keypoint	COCO keypoint (test-dev)	AP: 66.7, AP^M: 62.4, AP^L: 72.9
Depth Estimation	DPT-Hybrid (Ranftl et al., 2021)	MIX-6	DIW	WHDR: 11.06
Surface Normal	NLL-AngMF (Bae et al., 2021)	ScanNet	ScanNet (test)	Angular error (degrees) Mean: 11.8, Median: 5.7, RMSE: 20.0
Edge Detection	DexiNed-a (Poma et al., 2020)	BIPED	BIPED (test)	ODS: 0.859, OIS: 0.867, AP: 0.905
Object Detection	UniDet (Zhou et al., 2022)	COCO + Objects365 + OpenImages + Mapillary	COCO (test), OpenImages (test), Mapillary (test), Objects365 (valid)	mAP: 52.9 (COCO), 60.6 (OpenImages), 25.3 (Mapillary), 33.7 (Objects365)
OCR	CharNet (Xing et al., 2019)	SynthM	ICDAR 2015	Acc: 71.6 (sen), 74.2 (in-sen)
Semantic Segmentation	Mask2Former (Swin-L) (Cheng et al., 2022a)	ADE20k	ADE20K (val)	mIoU (s.s.): 56.1, mIoU (m.s.): 57.3
Panoptic Segmentation	Mask2Former (Swin-L) (Cheng et al., 2022a)	COCO panoptic (train2017)	COCO panoptic (val2017)	PQ: 57.8, PQ^{th}: 64.2, PQ^{st}: 48.1, AP^{th}_{pan}: 48.6, $mIoU_{pan}$: 67.4
Caption Generation	$Prismer_{LARGE}$ (Liu et al., 2023)	Pre-train: COCO Caption (Karpathy train) + Visual Genome + Conceptual Captions + SBU captions + Conceptual 12M Fine-tune: COCO Caption (Karpathy train)	COCO Caption (Karpathy test)	BLEU@4: 40.4, METEOR: 31.4, CIDEr: 136.5, SPICE: 24.4

Fig. 6. Example of output of models included in FRESCO.v1 implementation.

FRESCO-Score allows for an in-depth exploration of which aspects of the images are mostly different and hence affect the final score offering significant benefits in terms of interpretability. In fact, two pairs of images may have comparable final distances, but differ with respect to heterogeneous aspects. For instance, a pair of images, despite having different chromatic properties, may depict a similar figurative content, and thus present a distance akin to another pair that, while similar at the plastic level, have different figurative contents. A user may also choose to weight each component differently, so as to cluster images based on specific properties. These considerations are in general not possible with methods that return a single distance value between the representations of the two compared images in the feature space of a black-box pre-trained neural network. By considering all the analysis one by one, FRESCO-Score enables us to appreciate the difference among two images at different scales, delving into even the most intricate details, such as the direction of the gaze of each single person depicted in the image.

FRESCO-Score needs to aggregate properties that are associated with the whole image (e.g., chromatic categories, main topics, place) and single subjects or objects in the images (e.g., the characteristics, emotions, pose, and gaze of a specific person). While the former allows for a direct comparison of values computed at the image level, the latter requires a mapping strategy to associate each person/object of the first image to a comparable instance, if any, in the second one. For convenience, in the following we will refer to them as *image-level* and *object-level* measures. We exclude from FRESCO-Score intermediate maps (e.g., semantic maps or line maps) and body poses, that will be included in further development.

Specifically, given two images I_i and I_j, FRESCO extracts a series of measurements using the set of image-level and object-level functions $\mathcal{F}_t := \{f_t^k(\cdot)\}$, with $k = 1...K$, and object-level functions $\mathcal{G}_t := \{g_t^l(\cdot)\}$, with $l = 1...L$, and $t \in \{\texttt{plastic}, \texttt{figurative}, \texttt{enunciational}\}$, defined in Section 4.2. Each image is associated with the set of objects $\{o_i^m\}_{m=0}^M$ and $\{o_j^n\}_{n=0}^N$ detected in each image and whose positions in the pixel area are described by the respective centroids $\{c_i^m\}_{m=0}^M$ and $\{c_j^n\}_{n=0}^N$. FRESCO-Score first computes a matching function $m(\cdot,\cdot)$ that, given the two sets of centroids $\{c_i^m\}$ and $\{c_j^n\}$, returns a set of matched pairs $\{(\hat{c}_i^p, \hat{c}_j^p)\}_{p=0}^P$ such that a matching cost function is minimized. Then, FRESCO-Score is computed as:

$$S = \alpha S_{\texttt{pla}}(I_i, I_j) + \beta S_{\texttt{fig}}(I_i, I_j) + \gamma S_{\texttt{enu}}(I_i, I_j) \tag{1}$$

where α, β, γ are configurable parameters (set to 1 in the rest of this paper), and each S_t, with $t \in \{\texttt{plastic}, \texttt{figurative}, \texttt{enunciational}\}$, is computed by aggregating the pertinent subset of features as follows:

$$S_t(I_i, I_j) = \sum_{k=1}^{K} d^k(f_t^k(I_i), f_t^k(I_j)) + \sum_{l=1}^{L}\sum_{p=0}^{P} d^l(g_t^l(\hat{o}_i^p), g_t^l(\hat{o}_j^p)) \tag{2}$$

where $(\hat{o}_i^p, \hat{o}_j^p)$ are the pairs of matched objects, $\mathcal{F}_t := \{f_t^k(\cdot)\}$ and $\mathcal{G}_t := \{g_t^l(\cdot)\}$ are the subset of functions used to extract the measures at each level t, and d^k and d^l are the normalized distance functions computed on each couple of image- and instance-level measures, respectively. Prior to aggregating, distances can be scaled in the range [0, 1], as done in the remainder of this paper, or standardized using a mean and a standard deviation precomputed on a large dataset.

Image-level measures. Palette similarity is derived from the CIELAB color difference between two single colors obtained as the weighted average of the two palettes, following the single (homogeneous) color difference model proposed in Pan and Westland (2018).

RGB color histograms are compared using the Hellinger distance, which is related to the Bhattacharyya coefficient as follows:

$$BC(H_i, H_j) = \sum_{x \in \mathcal{X}} \sqrt{H_i(x) \cdot H_j(x)} \tag{3}$$

$$d(H_i, H_j) = \sqrt{1 - BC(H_i, H_j)} \tag{4}$$

where H_i and H_j are the histograms of the two images I_i and I_j. The final distance is obtained as the average of the distances computed across three channels. For scalar measures such as brightness, saturation, face-background ratio and background average depth, the absolute error is considered. Binary measures such as grayscale and indoor/outdoor evaluate to 1 if the corresponding value is the same in both images, 0 otherwise. Scene classification is compared using the cosine similarity on the confidence vectors returned by the model. Image tags are compared using the Jaccard index (Real and Vargas, 1996), while for the spatial coverage a continuous Jaccard index was properly designed, taking into account the common area for each category. The number of people and objects in the image are compared using the percentage of common instances. The caption is instead compared using the cosine similarity between the text embeddings extracted by the CLIP ViT/L-14 text encoder. All distances are scaled in the range [0,1].

Mapping strategy. In FRESCO.v1 each instance in the first image is associated with the closest instance of the same category in the second image using the centroids derived from the bounding boxes (faces and objects) or the instance masks (for analysis involving depth information).

The set of centroids $C_i := \{c_i^m\}_{m=0}^M$ identified in the first image I_i is associated with the set of centroids $C_j := \{c_j^n\}_{n=0}^N$ found in the second image I_j, minimizing the cost of matching. Specifically, given the two sets C_i and C_j, and a matching cost function $E : C_i \times C_j \to \mathbb{R}$, the objective is to find a bijection $f : C_i \to C_j$ such that the total cost of matching $\sum_{c_i \in C_i} E(c_i, f(c_i))$ is minimized. The cost E is defined as the squared Euclidean distance between each pair of centroids in the bipartite graph. To solve this problem, we leverage the SciPy's modified Jonker–Volgenant algorithm for linear sum assignment, which has a complexity of $O(n^3)$ in the worst case (Crouse, 2016).

Instance-level measures. All instances in the two images are compared one-by-one after executing the mapping algorithm. Object positions (vertical ratios, horizontal ratios, centralities, distances from cameras) are compared using the absolute error. Person characteristics are compared using the cosine similarity on the confidence vectors returned by the models for both multiattribute (i.e., 40 face attributes and gender)

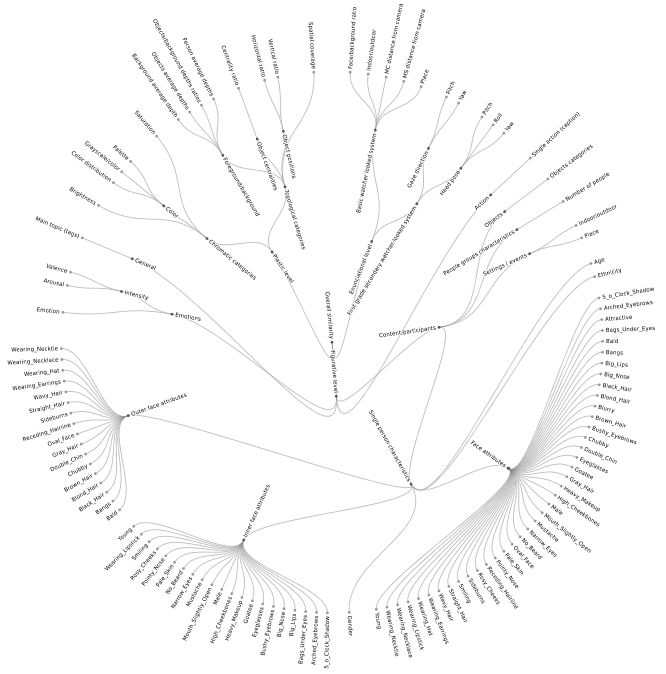


Fig. 7. The hierarchy of similarities on which Fresco Score is built. Best viewed online.

and multiclass (i.e., ethnicity and emotion) classifications. Continuous values such as age, valence, arousal and angles (roll, pitch, yaw) for both head pose and gaze direction are compared using the absolute error. Unpaired objects and faces (i.e., those for which the matching algorithm does not return a real association) are assigned by default to the minimum similarity value for all the associated measures. All distances are defined in or are scaled to the range [0,1].

Aggregation. In FRESCO.v1, all similarity values are aggregated using a linear weighted average. Instance-level similarities are combined to obtain an image-level value for each analysis. Specifically, the similarities computed for each characteristic of each individual face in the image are averaged to derive a single compound measurement. Likewise, similarities related to single objects were also averaged out. Similarity scores are further aggregated to obtain a cumulative measure for groups of correlated characteristics (e.g., chromatic, topological, etc.), and then each group of measures is further averaged to get a per-level measure. The final overall score is obtained by averaging the measures of the three levels of analysis (i.e., plastic, figurative, and enunciational). An overview of the hierarchy of similarity scores and their relative aggregations is illustrated in Fig. 7. It should be noted that, while all individual similarities are scaled between 0 and 1, their distribution may differ in practice. In the future, a more sophisticated aggregation will be considered, in which distance measures are calibrated on a target population.

6. Experimental validation

6.1. Datasets

Experimental validation was carried out on the FFHQ in-the-wild validation set (Karras et al., 2018) composed of 10,000 images and on the MIAP (More Inclusive Annotations for People) (Schumann et al., 2021) extension of the OpenImages v7 database (Benenson and Ferrari, 2022; Kuznetsova et al., 2020), which we further filtered by specifically considering the presence of at least one visible human face, obtaining a total of 2002 images. The former dataset is composed of uncropped original images scraped from Flickr. It encompasses a wide range of variations in terms of age, ethnicity, and image background. In addition, it also features a relevant assortment of accessories, including eyeglasses, hats, and more. The latter was instead selected because of its greater variability of contents, which includes more complex scenes. It is also equipped with additional annotations with respect to the original OpenImages including exhaustive bounding boxes for all people and attribute labels such as the human perceived gender and perceived age range.

6.2. Validation

There are two fundamental questions that must be kept in mind when validating a computational pipeline such as FRESCO, and concern its validity (i.e., how faithful is the information extracted from FRESCO to the original images) and usefulness (i.e., to what extent the measures extracted by FRESCO can be used to answer interesting research questions that complement and extend traditional manual extraction).

In this paper, we focus first and foremost on assessing the validity of the information extracted. Although each component included in FRESCO.v1 has been previously tested in isolation (relevant performance metrics are included in Table 2), some residual errors (both random and systematic) are unavoidable. These aspects are here investigated by searching for discrepancies and inconsistencies in the output of different tools, which not only provides an indirect measure of performance, but can also guide researchers from different fields in interpreting the results. The FRESCO-Score was validated by visually analyzing how it ranked images in terms of similarities.

7. Results

An example of application of the FRESCO.v1 computational pipeline is depicted in Fig. 8. It should be noted that many of the CV tools included in FRESCO.v1 produce a spatial output, such as a depth map or segmentation map, as exemplified in Fig. 6 and in the left part of Fig. 8. Keypoints and spatial maps are unstructured data and are difficult to analyze at scale; for this reason, FRESCO.v1 converts them into a set of interpretable indicators, either numerical or categorical, that lend themselves to analytics techniques. An example is provided in the right part of Fig. 8, in which quantities are divided according to whether they are pertinent to the plastic, figurative, or enunciational level.

In the rest of this Section, we first evaluate the consistency of models (Section 7.1) and then proceed to evaluate the FRESCO-score similarity measure (Section 7.2).

7.1. Accuracy and consistency of the extracted quantities

First, we seek to answer questions related to how the output of different measures can be combined and compared. For instance, in Fig. 8 it is evident that the estimated age is incorrect (25 years). However, both the caption and the tags refer to the presence of a young boy or a child, which is an indication that the age may not be accurately estimated. Two persons and two horses are detected by semantic segmentation and object detection, but only one pose and face are detected: indeed, an arm and the back of a horse are visible in the bottom left corner. An uncommon object (a red tank) is misclassified by the caption as a fire idrant, but the latter is not identified by the object detector, nor the semantic segmentation. A more systematic analysis of these discrepancies is presented in Tables 3 and 5.

Determining whether two outputs are consistent is not straightforward as FRESCO involves multiple models whose set of labels may differ substantially, requiring some form of concept mapping. Some concepts may be expressed only by one label set, or may be expressed by different label sets at varying granularities. The problem is more evident when models with a predefined, closed-set output label space, such as object detectors, are compared with open-set models such as tagging or captioning models. As an example, the caption generator can produce

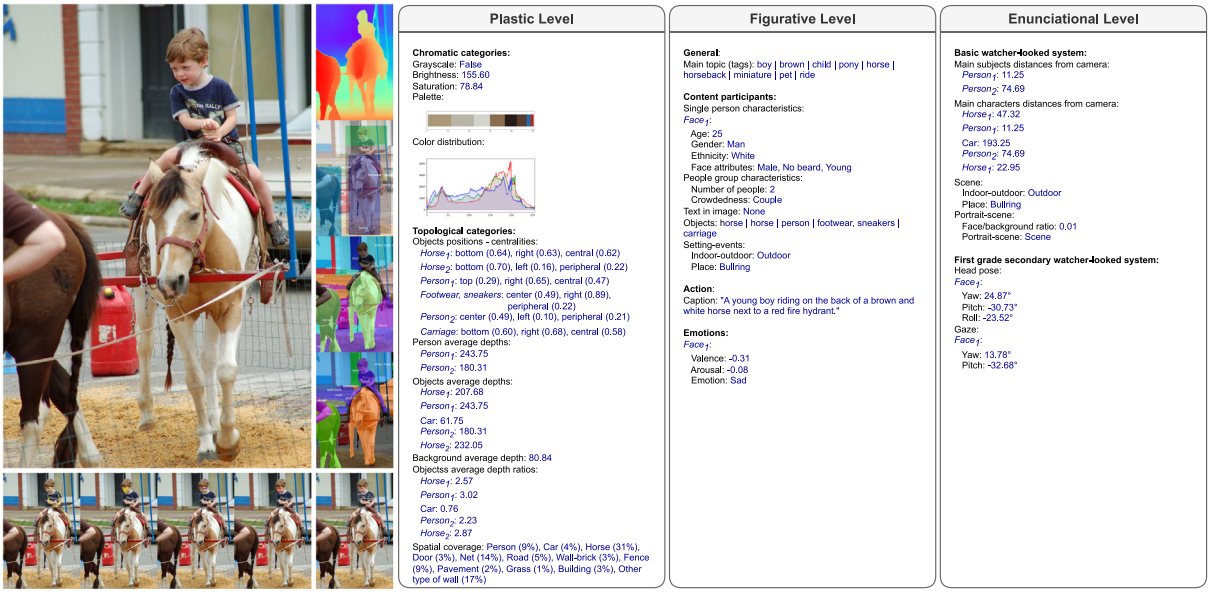


Fig. 8. Example of FRESCO.v1 final output. All quantities extracted are defined in Table 1. The figurative and plastic level are closely intertwined. Notice, for instance, how the figurative content is that of a boy riding a horse, but the spatial disposition of the figures is part of the plastic level (the horse is positioned in the center and occupies a substantial portion of figure, and that the boy is located in the top right part of the image).

Table 3
Analysis of the people detection consistency across different models included in FRESCO.v1 evaluated on OpenImages and FFHQ in-the-wild datasets.

	People detected (OpenImages)	
Tasks	Images with people (at least one)	People per image (all detected)
Face detection	90.26%	2.07
Object detection	88.31%	1.85
Panoptic segmentation	92.01%	2.14
Semantic segmentation	95.55%	–
Tagging	96.10%	–
Captioning	95.35%	–
	People detected (FFHQ in-the-wild)	
Tasks	Images with people (at least one)	People per image (all detected)
Face detection	100.00%	1.75
Object detection	99.21%	2.03
Panoptic segmentation	99.84%	2.27
Semantic segmentation	99.97%	–
Tagging	99.99%	–
Captioning	99.98%	–

outputs in the form *"A baseball player catching a ball…"*, *"A family posing for a picture…"*, *"A mother and daughter pose for a picture…"* that suggest the presence of people in the image even if the concept "person" is not explicitly mentioned. On the other hand, the object detector used in Fresco.v1 does not consider label hierarchies (Zhou et al., 2022), hence the class "person" is considered as an independent concept with respect to "man", "woman", "girl" and "boy" which are also included in the label space.

Table 3 reports *the number of people (2.2.1.1) detected by each model or task*. To achieve a more reliable estimate of the number of people detected by each model, a broad synset was properly selected to represent the concept "person" including all labels and expressions that can be traced back to the original wider concept. The results show that the number of detected faces is lower than the number of people found by the object detector and the panoptic segmentation, especially on FFHQ in-the-wild. This should not necessarily be interpreted as an indication of poorer performance of the face detector, but, more likely, this result may be attributable to the presence of people photographed from behind or with occluded/cut faces. Unexpectedly, the panoptic model seems to retrieve a slightly larger number of persons. It should be noted that, unlike object detection, the panoptic label space and semantic segmentation include both "things" and "stuff" categories. Consequently, they are able to catch information about both countable objects which are characterized by a well-defined shape (things) and uncountable categories which are in general amorphous and belong predominantly to the context of the scene (stuff). Hence, this difference may be explained by the huge gap in the number of categories taken into account by the two models, only 133 (80 "things" + 53 "stuff") compared to 722. The percentage of images recognized as containing people is close to 100% for all models on FFHQ in-the-wild. Instead, despite the OpenImages split in use being properly filtered in the presence of "Human face", the percentage is in general lower. Unlike FFHQ in-the-wild, in OpenImages a limited number of sketches and cartoons are included, since they were annotated as containing "Human face"; however, models such as the face detector are trained on real faces and may fail in these different domains. The models used for tagging and captioning appear to be more robust also in these images. Last, in Table 4 we report the distribution of the number of people discretized according to the categories defined in Table 1 (that is, no people, single person, couple, small group, medium group, large group or crowd).

The consistency among the topics (2.1.1) and the objects (2.2.3) detected by each model was further evaluated on the OpenImages validation set (Table 5). To make semantic similar labels comparable, we leveraged a variant of the CLIP score which evaluates the cosine similarity between the text embeddings of the two labels extracted using the CLIP ViT-L/14 text encoder. To establish whether a concept is equally recognized by different models, we set a threshold on the similarity score. Some labels, despite referring to the same concept, may use different words and/or include more details. For example, labels "land vehicle", "sport car", and "sedan" are more specific cases of the general concept "car"; compared to label "car", their CLIP score is 0.83. Depending on the threshold selected, we may consider semantically related information as equivalent or not. We compared each pair of tasks (e.g., image tagging vs. object detection, image tagging vs. panoptic segmentation) to determine whether on average each task provides more, equal, or less information than the other (i.e., whether the output contains the same concept, or whether certain concepts are present only in one of the outputs), at a given CLIP

Table 4
Analysis of the people groups detection consistency across different models included in FRESCO.v1 evaluated on OpenImages and FFHQ in-the-wild datasets. In FRESCO.v1 counts are inferred from the object detector to adjust for individuals seen from behind.

	Images with groups of people (OpenImages)					
Tasks	0/1 (single)	2 (couple)	3–6 (small group)	7–12 (medium group)	13–30 (large group)	31+ (crowd)
Face detection	67.88%	15.78%	11.94%	2.85%	1.25%	0.30%
Object detection	60.84%	17.38%	18.63%	3.00%	0.15%	0.00%
Panoptic segmentation	61.04%	16.28%	16.18%	5.54%	0.95%	0.00%
	Images with groups of people (FFHQ in-the-wild)					
Tasks	0/1 (single)	2 (couple)	3–6 (small group)	7–12 (medium group)	13–30 (large group)	31+ (crowd)
Face detection	61.73%	21.80%	15.06%	1.17%	0.24%	0.00%
Object detection	52.70%	21.80%	23.56%	1.94%	0.00%	0.00%
Panoptic segmentation	51.99%	21.79%	20.60%	5.40%	0.22%	0.00%

Table 5
Analysis of the topics detection consistency across different models included in FRESCO.v1 evaluated on OpenImages dataset. It considers three different thresholds on the labels encoding similarities to establish if a topic can be considered in common among the predictions of each pair of models.

Tasks	Topics detected (OpenImages)								
	In first	In common ($CLIP_{Score} \geq 0.80$)	In second	In first	In common ($CLIP_{Score} \geq 0.85$)	In second	In first	In common ($CLIP_{Score} \geq 0.90$)	In second
Tags-Objects	60.12%	29.86%	10.01%	71.91%	11.58%	16.52%	75.16%	5.47%	19.37%
Tags-Semantic	50.37%	30.09%	19.53%	64.40%	9.23%	26.37%	67.00%	4.63%	28.36%
Tags-Panoptic (things)	72.03%	25.70%	2.27%	86.22%	7.82%	5.96%	88.21%	3.48%	8.31%
Objects-Semantic	23.16%	26.30%	50.54%	30.29%	16.21%	53.49%	30.83%	15.21%	53.97%
Objects-Panoptic (things)	47.94%	43.38%	8.68%	62.09%	27.96%	9.95%	62.69%	26.87%	10.43%
Semantic-Panoptic (things)	65.49%	34.51%	0.00%	71.84%	28.16%	0.00%	72.27%	27.73%	0.00%

score threshold. The results of Table 5 indicate that image tagging can associate the highest number of topics with a given image. Compared to image tagging, semantic segmentation can identify the highest number of additional topics (around 20% at a threshold of 0.8), followed by the object detector with about 10%. Semantic segmentation grasps more concepts w.r.t. to object detection as it embraces both labels from countable objects (things) and uncountable categories (stuff) which characterize mainly the background. This is further supported by the results achieved by the semantic segmentation, which adds more than 50% topics to the object detector. In turn, the object detector is able to find much more topics compared to the panoptic segmentation (things) due to its higher label space (722 vs. 80). Panoptic segmentation cannot detect more topics than semantic segmentation, as it was trained on the same dataset and its label space is a subset of the latter. In this case, the topics in common are likely to be the things identified by both models. Lastly, the use of multiple models trained on different datasets introduces a relevant benefit: it allows to capture a wide range of information from the image compensating any oversight of each single model. In fact, common topics are less likely to arise from mispredictions of individual models.

Finally, we investigated *the distributions of a subset of continuous and categorical measures* extracted from the FFQH-in-the-wild and OpenImages validation set (Figs. 9 and 10). In terms of plastic categories, the centroids of objects and persons appeared to be predominantly located *in the middle of* the picture frame both horizontally (1.3.2) and vertically (1.3.1). Both datasets have similar characteristics in terms of *brightness* (1.2.2) and *saturation* (1.2.3).

At the figurative level, all images depict close-up portraits or scenes in which two or more people interact. Unlikely OpenImages, FFHQ in-the-wild shows a bimodal distribution for the *Valence* (2.5.3) category, which is consistent with a substantial presence of people smiling and posing for the camera. *Emotion classification* (2.5.2) further supports this finding, since the FFQH in-the-wild distribution reaches its peak in the "happy" category, while for OpenImages the dominant class is "neutral". The *age* (2.2.2.2) distribution is quite similar for both datasets varying mostly in the range 20–50 (age is normalized between 0 and 100). In terms of *ethnicity* (2.2.2.5), both datasets are imbalanced with a strong prevalence of "white" and "asian" categories, while others are markedly underrepresented.

At the enunciational level, with a few exceptions (more evident in OpenImages), the gaze (3.2.3) and head (3.2.1) angles peaked around the value of 0.5, indicating the predominant presence of people looking at the camera while posing for pictures. The face/background ratio (3.1.3) is skewed in the 0–0.2 range, even more evident for the OpenImages split, suggesting that the majority of images are scenes depicting people in context, rather than portraits. The current implementation sets a threshold at 0.3, so an image is considered a portrait if the largest face box covers at least 30% of the total image area.

The *main subject's distance from camera* (3.1.1) is a bimodal distribution, suggesting the presence of two main groups: close-up portraits and scenes in which several persons interact.

7.2. Validation of the FRESCO score

The FRESCO Score can be employed to rank images based on their similarity. Images can be compared using measures at different levels of aggregation enabling comparisons at varying degrees of detail, from the more general (Overall Score) to specific groups of aspects pertaining to the three levels of analysis (Plastic, Figurative and Enunciational Score) or even down to the more fine-grained characteristics, such as expressions, ethnicity, and head/gaze orientation of each single person depicted in the image, through measures at the lowest level of the FRESCO Score hierarchy. Fig. 11 shows an example of ranking using the Score at the highest level and the three main levels of analysis. In this case, the reference image is compared with the entire FFHQ in-the-wild validation set consisting of 10,000 images. The retrieved groups of images highlight that the Plastic Score, which includes Chromatic (1.2) and Topological categories (1.3), is more susceptible to colors variations (the more distant samples are in general darker) and spatial dispositions of forms (both in terms of covered pixels and distance from camera). The Figurative Score, which covers analysis on characteristics of each person (2.2.2) including among others emotions, gender, and face attributes, allows to retrieve images of people with similar characteristics. In fact, all the images in the Top-8 contain two women of comparable age posing in an outdoor environment, and the presence

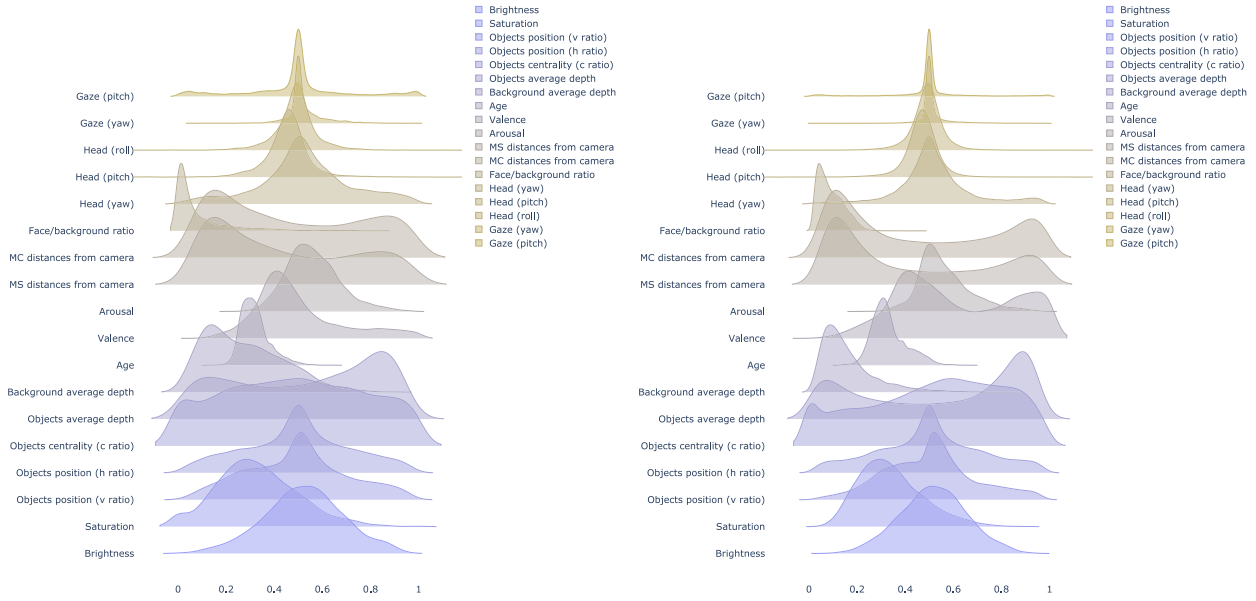


Fig. 9. Distribution of a selection of numerical values in the OpenImages (left) and FFHQ-in-the-wild (right) subsets.

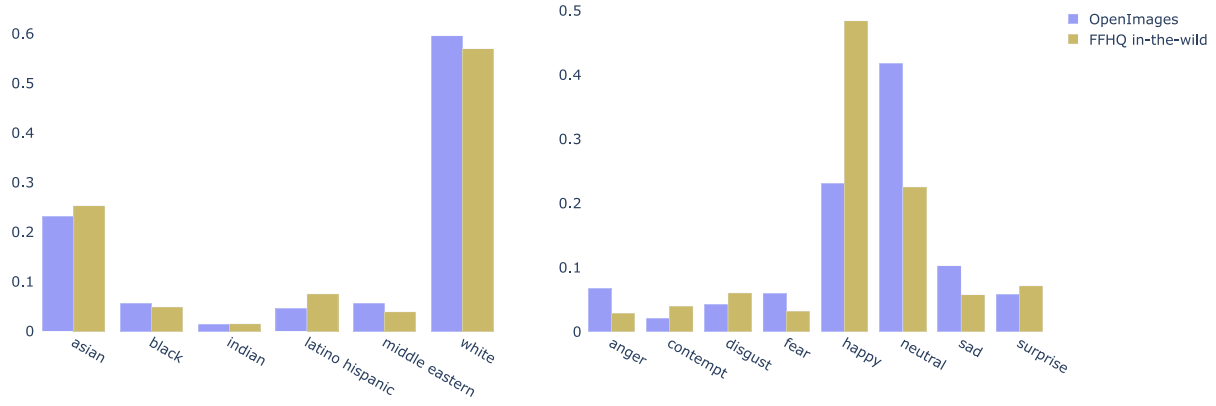


Fig. 10. Distribution of a selection of categorical measures in the OpenImages and FFHQ-in-the-wild subsets.

of accessories such as hats, glasses and necklaces is generally more consistent. The Enunciational Score is instead sensible to the mean distance from camera (3.1.1), the scene (3.1.3) and the direction of head/gaze (3.2.1, 3.2.3) (regardless of the precise spatial position of the subject) and other aspects such as gender, accessories, and so forth. The Overall Score combines the previous scores; when all scores are equally weighted, it can grasp the general content, but it can lose sensitivity to specific aspects. The weight of each level can be properly adjusted to emphasize certain characteristics in the retrieved images, depending on the interests and goals of the scholar using the platform. The number of people (2.2.1.1) is in general well preserved by all scores, due to the effect of the matching strategy that penalizes the presence of unpaired objects as stated in Section 5. Indeed, in all cases, the Last-8 images represent in general crowded scenes. The same considerations are also valid for images of a single person as shown in Fig. 12.

The ranking based on a subset of single analysis is illustrated in Fig. 13. Each score allows us to find images that are comparable in that specific aspect. Given a reference image, each row shows images that are closest (or farthest) in terms of Color distribution (1.2.1), Textual description (2.4.1), Spatial coverage (1.3.5), and General topics (2.1.1), and in terms of characteristics of single faces including Ethnicity (2.2.2.5), Emotion (2.5.2), Head pose (3.2.1), and Gaze direction (3.2.3). For this specific test, the unpaired objects were excluded from analysis, hence the comparison on faces involves only those who can be directly matched to a comparable one in the second image. Working directly with distances on the models outputs, the Top-8 and the Last-8 images are exactly at the opposite for each specific analysis (within the limit of variability covered by the FFHQ-in-the-wild Validation Set) with the Median-8 in the middle of the distribution, as is definitely evident for head and gaze angles.

8. Discussion

In this work, we thoroughly explore the application of structural visual semiotics principles to develop a detailed computational framework that facilitates the analysis of large-scale image archives. In this way, semioticians, as well as scholars in the social sciences and humanities in general, can leverage recent advances in computer vision, and particularly the availability of general purpose models pre-trained on vast amounts of data, also known as foundation models (Zhou et al., 2023). In the context of social media, for instance, FRESCO can be used to answer questions such as: when self-representing themselves, do people want to show themselves happy or do they prefer to show other moods? Is their face the focus of their images? Do people usually

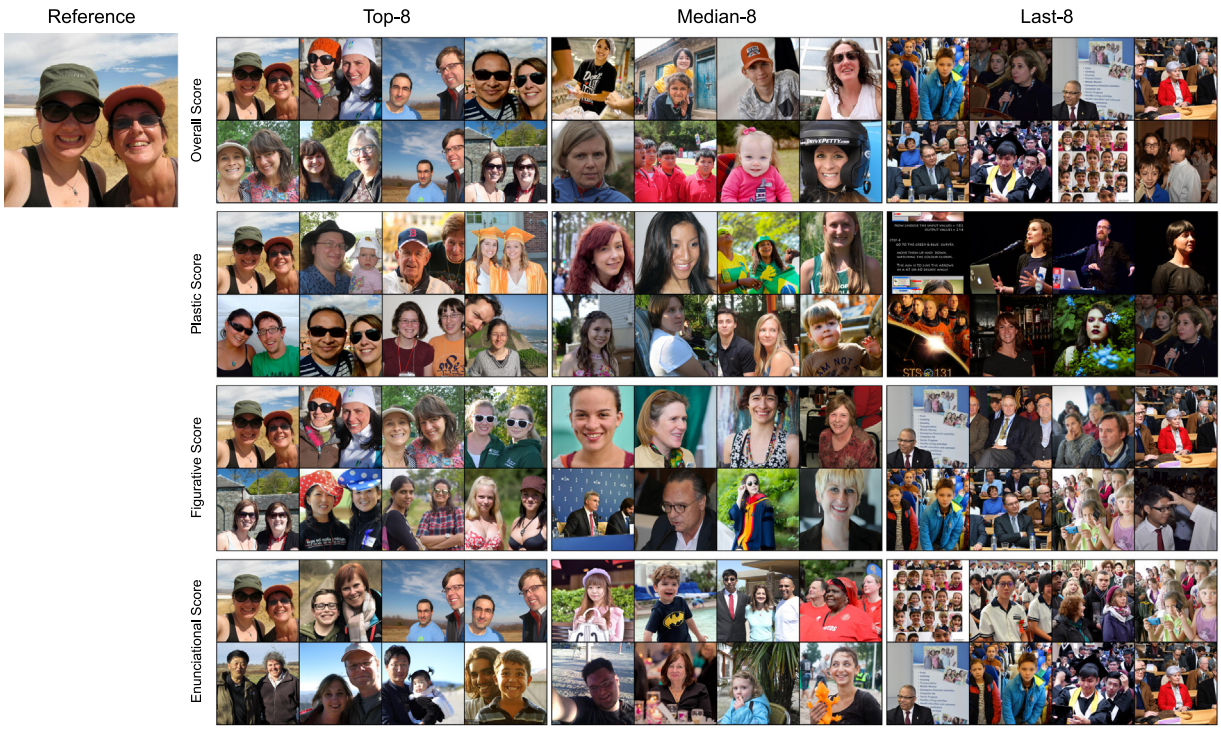


Fig. 11. Ranked images using the three scores derived from the highest levels of analysis and the overall score on a reference image including multiple people. Each group of images (Top-8) shows different common aspects such as colors and spatial dispositions of forms (plastic score), person characteristics such as age, gender, emotions, accessories (figurative score) and distances from camera, head/gaze directions while not caring about other details such as gender, emotions an so forth (enunciational score).

Fig. 12. Ranked images using the three scores derived from the highest levels of analysis and the overall score on a reference image including a single person. Each group of images (Top-8) shows different common aspects such as colors and spatial dispositions of forms (plastic score), person characteristics such as age, gender, emotions, accessories (figurative score) and distances from camera, head/gaze directions while not caring about other details such as gender, emotions an so forth (enunciational score).

show themselves alone or in company? and many other questions as discussed in Santangelo and Morra (in press).

The outcome of our research is a computational platform that can serve several purposes. First, it converts an image collection from unstructured image data to structured data in tabular format to support the application of data analytics tools (Santangelo and Morra, in press). This tabular format summarizes to what extent the different traits commonly employed by semioticians to characterize images are

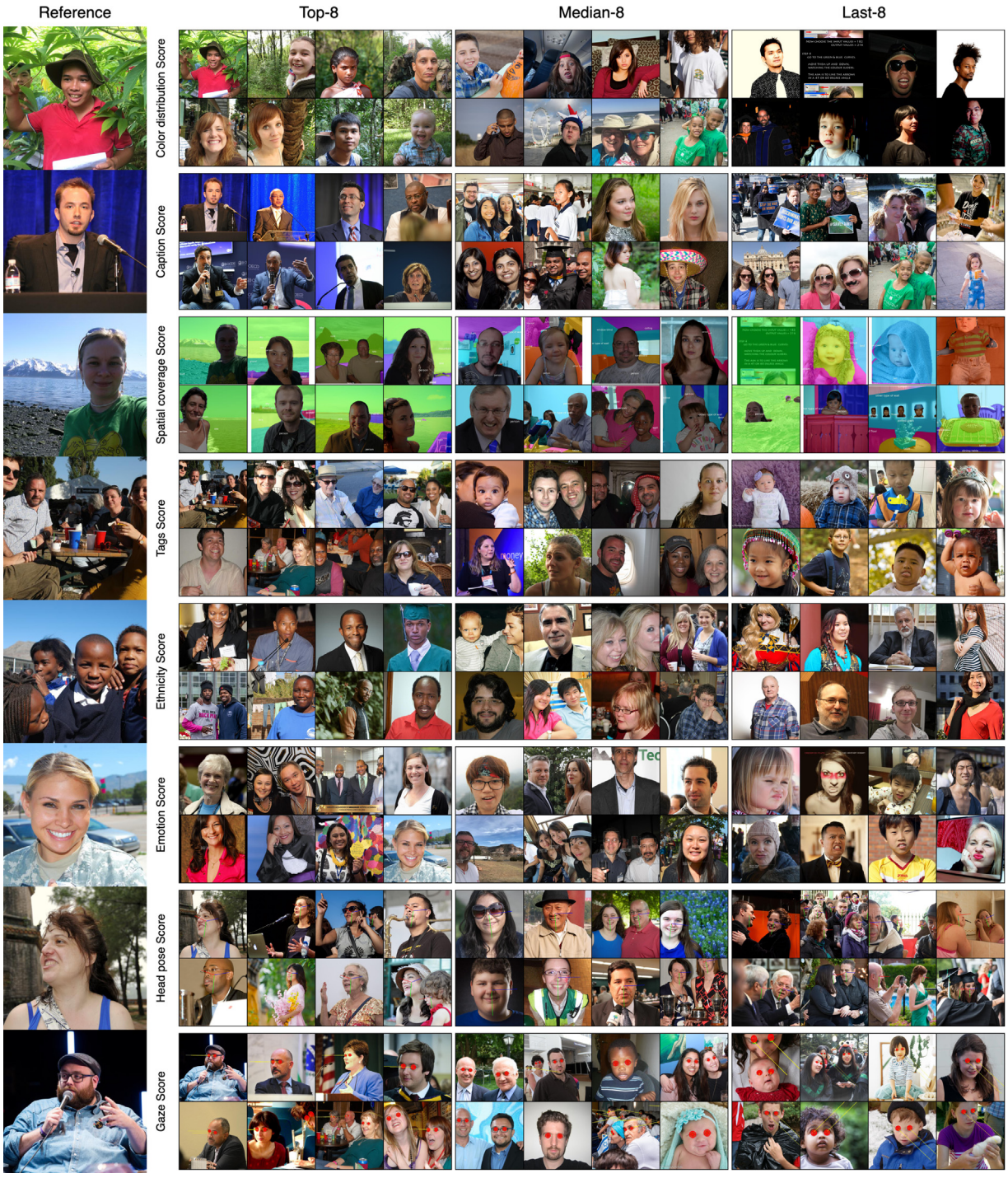


Fig. 13. Ranked images using the scores derived from the lowest levels of analysis on a various set of reference images including single and multiple people. Each group of images (Top-8) shows different common aspects strictly related to the considered analysis. For a better interpretability of the results, in this specific test unpaired objects were excluded from the analysis, hence all scores referring to face aspects consider only faces for which a comparable instance can be found among both images, neglecting all the others.

expressed by each image, thus configuring a sort of*digital identikit* of an image. Second, it supports a content-based image retrieval system that is based on the plastic, figurative, and/or enunciational content of images, using a configurable similarity score (FRESCO-Score). That said, it is also possible to search for similar images (content-based image retrieval) or to group images (clustering) based on individual characteristics such as caption, human pose, spatial composition, or color distribution. This enables the discovery of similarities and groupings in the data that might otherwise remain unnoticed by researchers (Männistö et al., 2022). Lastly, FRESCO could be used to analyze the quality and content of synthetic images produced by generative models, e.g., to measure alignment with specific instructions, or to quantify systematic differences introduced by generative models with respect to natural images (Piano et al., 2024b,a; Barattin et al., 2023; Otani et al., 2023).

The experimental validation in Section 7, along with the standalone performance in Table 2, highlights how the extracted features are, in general, of high quality. The agreement between models is generally good, and the distributions extracted from the two datasets analyzed are consistent with how they were sourced and collected. However,

there are still challenges and limitations associated with the current implementation of the FRESCO pipeline. Computer vision techniques, although increasingly accurate, may inject various types of errors. Algorithms included in the FRESCO pipeline were selected based on their performance, but noise in the form of errors or uncertainties can arise due to factors such as variations in lighting conditions, image quality, or the complexity of the subject matter. Agreement between different models can be used to filter out uncertain and possible erroneous output. Care must be taken when comparing findings across datasets, but as long as errors can be expected to occur at approximately the same rate on each dataset, a relative comparison can be more reliably estimated than an absolute value (Männistö et al., 2022).

Another challenge that arises when using deep neural networks, especially those relying on a closed set of labels, are out-of-distribution samples. Deep learning models targeting the human face and body, such as for keypoint detection or classification of facial expression, are less prone to this drawback; on the other hand, networks that perform complex and potentially ambiguous multi-label classification, such as scene classification, should be interpreted with greater caution. In the future, we plan to expand the FRESCO implementation with out-of-distribution detection (Recalcati et al., 2023; Yang et al., 2024; Fort et al., 2021) or one-shot domain adaption (Yang et al., 2024; D'Innocente et al., 2020).

Besides such practical issues, there are also a few limitations that derive from the current architecture of FRESCO. In the field of structural visual semiotics, on which FRESCO foundations were established, the meaning of an image depends not only on what is present and can be seen but also on what it is omitted. Just as we are able to understand the meaning of a sign such as "l" (the letter "L" of the alphabet) as different from "t", because essentially in the long vertical bar we see there is a short horizontal one at the top missing (De Saussure, 1989), in the same way, we understand that the meaning of a censored image is that subjects seen in other images are not represented. If we look at the famous painting by Manet's "Olympia", we see that in her nudity she looks at us proudly, directly and from above, instead of from below, with a more demure and indirect gaze, surrounded by a maid and a cat, instead of a governess and a dog, as in Titian's painting titled "Venus of Urbino". Manet's work evidently wants to differentiate itself from Titian's one and from a certain tradition in the representation of naked women who are aware to be observed from men (Berger, 2008), hence we realize that its significance depends, precisely, on the absence of some very significant elements of Titian's own painting and the presence of other deliberately different ones. To overcome this critical gap, FRESCO should be extended with the ability to select, attend, and reason about external and commonsense knowledge (Joshi et al., 2024; Ye and Kovashka, 2018).

9. Conclusions

In this study, we extensively investigate the use of structural visual semiotics principles to create FRESCO, a comprehensive computational framework that supports scholars in the analysis of large-scale image archives by leveraging the power of foundational models. In constructing FRESCO, instead of deploying a makeshift collection of deep neural networks, we aimed to represent each category of semiotics through numerical values that can be derived using cutting-edge computer vision models.

We expect FRESCO to further promote the adoption of quantitative methods in visual semiotics (Manovich, 2020), closing the gap with respect to the analysis of text corpora. At the same time, we hope that FRESCO can foster the interdisciplinary collaboration between computer vision scientists and humanities scholars (Bocyte and van Kemenade, 2022).

The present study has focused on the technical characteristics of the FRESCO pipeline, outlining its design principle compared to previous studies (Männistö et al., 2022). We also performed internal validation with the aim of investigating the consistency and usability of different extracted characteristics. Currently, we are planning to apply FRESCO to selected case studies involving real-life image collections. In future studies, our aim is to further expand FRESCO by expanding the set of characteristics computed. FRESCO should also be extended to identify and connect elements that are found in an image with those that are absent, but are nonetheless connected to it. To this aim, FRESCO should be integrated with the ability to integrate external and commonsense knowledge, either in the form of structured Knowledge Graphs and/or embedded in Multimodal Large Language Models. Finally, further directions include making the computational pipeline more robust, including out-of-distribution detection, as well as adapting and validating the pipeline in other types of image archives, such as historical photography, advertisements, and paintings.

CRediT authorship contribution statement

Lia Morra: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Antonio Santangelo:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Pietro Basci:** Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Luca Piano:** Writing – original draft, Visualization, Software, Data curation. **Fabio Garcea:** Software, Methodology. **Fabrizio Lamberti:** Supervision, Methodology. **Massimo Leone:** Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this research are publicly available. A link to the code repository is included in the manuscript.

Acknowledgments

The present research is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No 819649-FACETS; PI: Massimo LEONE). The authors wish to thank Pietro Recalcati, Marco Porro, and Enrico Clemente for contributing to the system development. The authors also thank the entire FACETS research team for the stimulating discussions.

References

Amirshahi, S.A., Hayn-Leichsenring, G.U., Denzler, J., Redies, C., 2014. Evaluating the rule of thirds in photographs and paintings. Art Percept. 2, 163–182.

Arnold, T., Tilton, L., 2019. Distant viewing: analyzing large visual corpora. Dig. Scholarship Humanities 34, i3–i16.

Arnold, T., Tilton, L., 2020. Enriching historic photography with structured data using image region segmentation. In: Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access. pp. 1–10.

Arnold, T., Tilton, L., Wigard, J., 2022. Automatic identification and classification of portraits in a corpus of historical photographs. p. 0073, Proceedings http://ceur-ws.orgISSN, Vol. 1613.

Bae, G., Budvytis, I., Cipolla, R., 2021. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: International Conference on Computer Vision. ICCV.

Barattin, S., Tzelepis, C., Patras, I., Sebe, N., 2023. Attribute-preserving face dataset anonymization via latent code optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8001–8010.

Bell, P., 2012. Content analysis of visual images. SAGE Visual Methods: Interpret. Classif. 3, 31–57.

Benenson, R., Ferrari, V., 2022. From colouring-in to pointillism: revisiting semantic segmentation supervision. arXiv preprint arXiv:2210.14142.
Berger, J., 2008. Ways of Seeing. Penguin UK.
Berlanga-Fernández, I., Reyes, E., 2024. The digital approach to semiotics: a systematic review. Text Talk 44, 119–140.
Bianconi, F., Fernández, A., Smeraldi, F., Pascoletti, G., 2021. Colour and texture descriptors for visual recognition: A historical overview. J. Imaging 7 (245).
Bocyte, R., van Kemenade, P., 2022. AI Techniques and Tools for Social Sciences and Humanities Research. White Paper. Technical Report, NISV - Netherlands Institute for Sound & Vision.
Branz, L., Brockmann, P., Hinze, A., 2020. Red is open-minded, blue is conscientious: Predicting user traits from instagram image data. In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. pp. 23–28.
Chen, Q., Carneiro, G., 2015. Artistic image analysis using the composition of human figures. In: Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12 2014, Proceedings, Part I 13. Springer, pp. 117–132.
Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022a. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.
Cheng, J., Wang, L., Wu, J., Hu, X., Jeon, G., Tao, D., Zhou, M., 2022b. Visual relationship detection: A survey. IEEE Trans. Cybern. 52, 8453–8466.
Cholet, C., 2018. Images as utterances and as multimodal perceptual experiences. Quant. Semiotic Anal. 10, 1–120.
Compagno, D., 2018. Quantitative Semiotic Analysis. Springer.
Corrain, L., Valenti, M., 2023. Leggere L'Opera D'Arte. Dal Figurativo All'Astratto. Società Editrice Esculapio.
Cristani, M., Del Bue, A., Murino, V., Setti, F., Vinciarelli, A., 2020. The visual social distancing problem. Ieee Access 8, 126876–126886.
Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V., 2011. Towards computational proxemics: Inferring social relations from interpersonal distances. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE, pp. 290–297.
Crouse, D.F., 2016. On implementing 2D rectangular assignment algorithms. IEEE Trans. Aerosp. Electron. Syst. 52, 1679–1696.
Cucurull, G., Rodríguez, P., Yazici, V.O., Gonfaus, J.M., Roca, F.X., Gonzàlez, J., 2018. Deep inference of personality traits by integrating image and word use in social networks. arXiv preprint arXiv:1802.06757.
Cutzu, F., Hammoud, R., Leykin, A., 2003. Estimating the photorealism of images: Distinguishing paintings from photographs. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. IEEE, pp. II–305.
Datta, R., Joshi, D., Li, J., Wang, J.Z., 2006. Studying aesthetics in photographic images using a computational approach. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May (2006) 7-13, Proceedings, Part III 9. Springer, pp. 288–301.
De Saussure, F., 1989. Cours de linguistique générale. Vol. 1. Otto Harrassowitz Verlag.
Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S., 2020. RetinaFace: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5203–5212.
D'Innocente, A., Borlino, F.C., Bucci, S., Caputo, B., Tommasi, T., 2020. One-shot unsupervised cross-domain detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August (2020) 23–28, Proceedings, Part XVI 16. Springer, pp. 732–748.
Dondero, M.G., 2020. I linguaggi dell'immagine: Dalla pittura ai big visual data. Mimesis.
Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., Mazzone, M., 2018. The shape of art history in the eyes of the machine. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Eugeni, R., 2014. Analisi Semiotica Dell'Immagine: Pittura, Illustrazione, Fotografia. EDUCatt-Ente per il diritto allo studio universitario dell'Università Cattolica.
Ferwerda, B., Schedl, M., Tkalcic, M., 2015. Predicting personality traits with instagram pictures. In: Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015. pp. 7–10.
Fort, S., Ren, J., Lakshminarayanan, B., 2021. Exploring the limits of out-of-distribution detection. Adv. Neural Inf. Process. Syst. 34, 7068–7081.
Ghidoli, D., Montanari, F., 2021. Finding a socio-semiotic role for data science: A review of applications and case studies and some critical reflections. Mediascapes J. 7, 5–103.
Greimas, J.A., 1984. Sémiotique figurative et sémiotique plastique. Actes sémiotiques 24.
Hall, E.T., 1966. The hidden dimension. garden city, nueva york.
Hempel, T., Abdelrahman, A.A., Al-Hamadi, A., 2022. 6D rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 2496–2500.
Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y., 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718.
Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a nash equilibrium. CoRR abs/1706.08500.
Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L., 2022. Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17980–17989.
Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L., 2023. Inject semantic concepts into image tagging for open-set recognition. arXiv preprint arXiv:2310.15200.
Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A., 2017. Automatic understanding of image and video advertisements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1705–1715.
Joshi, S., Ilievski, F., Luceri, L., 2024. Contextualizing internet memes across social media platforms. In: Companion Proceedings of the ACM on Web Conference 2024. pp. 1831–1840.
Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations.
Kreiss, S., Bertoni, L., Alahi, A., 2019. PifPaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11977–11986.
Kress, G., Leeuwen, T.Van., 1996. Reading Images: The Grammar of Visual Design, first ed. Routledge.
Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int. J. Computer Vision 128, 1956–1981.
Liu, W., Chen, C., Wong, K.Y., 2018. Char-net: A character-aware neural network for distorted scene text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., Anandkumar, A., 2023. Prismer: A vision-language model with an ensemble of experts. arXiv preprint arXiv:2303.02506.
Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision. ICCV.
Lu, X., Sawant, N., Newman, M.G., Adams, R.B., Wang, J.Z., Li, J., 2016. Identifying emotions aroused from paintings. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and (2016) 15-16, Proceedings, Part I 14. Springer, pp. 48–63.
Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., et al., 2019. Mediapipe: A framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR At IEEE Computer Vision and Pattern Recognition. CVPR.
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. In: European Conference on Computer Vision. Springer, pp. 109–125.
Mangano, D., et al., 2018. Che cos' è la semiotica della fotografia.. Carocci.
Männistö, A., Seker, M., Iosifidis, A., Raitoharju, J., 2022. Automatic image content extraction: Operationalizing machine learning in humanistic photographic studies of large visual archives. arXiv preprint arXiv:2204.02149.
Manovich, L., 2020. Cultural Analytics. Mit Press.
Martinez Pandiani, D.S., 2024. The wicked problem of naming the intangible: Abstract concepts, binary thinking, and computer vision labels. Future Humanit. 2, e11.
Moretti, F., 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. Verso.
O'Halloran, K.L., 2015. Multimodal digital humanities. In: International handbook of semiotics. pp. 389–415.
Ortiz-Ospina, E., 2019. The rise of social media. Our World in Data https://ourworldindata.org/rise-of-social-media.
Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S., 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14277–14286.
Owyong, Y.S.M., 2009. Clothing semiotics and the social construction of power relations. Soc. Semiotics 19, 191–211.
Pan, Q., Westland, S., 2018. Comparative evaluation of color differences between color palettes. In: Color and Imaging Conference. Society for Imaging Science and Technology, pp. 110–115.
Pandiani, D.S.M., Presutti, V., 2023. Seeing the intangible: Surveying automatic high-level visual understanding from still images. arXiv preprint arXiv:2308.10562.
Pezzini, I., Spaziante, L., 2014. Corpi mediali. Semiotica e contemporaneità.
Piano, L., Basci, P., Lamberti, F., Morra, L., 2024a. Harnessing foundation models for image anonymization. In: 2024 IEEE Gaming, Entertainment, and Media Conference. GEM, IEEE, pp. 1–5.
Piano, L., Basci, P., Lamberti, F., Morra, L., 2024b. Latent diffusion models for attribute-preserving image anonymization. arXiv preprint arXiv:2403.14790.
Polidoro, P., 2008. Che cos' è la semiotica visiva. Carocci.
Poma, X.S., Riba, E., Sappa, A., 2020. Dense extreme inception network: Towards a robust cnn model for edge detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1923–1932.

Ramtoula, B., Gadd, M., Newman, P., De Martini, D., 2023. Visual DNA: Representing and comparing images using distributions of neuron activations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11113–11123.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188.

Real, R., Vargas, J.M., 1996. The probabilistic basis of jaccard's index of similarity. Systemat. Biol. 45, 380–385.

Recalcati, P., Garcea, F., Piano, L., Lamberti, F., Morra, L., 2023. Toward a realistic benchmark for out-of-distribution detection. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 1–10.

Reyes, E., Sonesson, G., 2019. New approaches to plastic language: Prolegomena to a computer-aided approach to pictorial semiotics. Semiotica 2019, 71–95.

Santangelo, A., Morra, L., 2024. FACE IT! the new challenges of visual semiotics. In: Chapter What Artificial Intelligence Tells Us About Ourselves. Routledge, (in press).

Santos, I., Castro, L., Rodriguez-Fernandez, N., Torrente-Patino, A., Carballal, A., 2021. Artificial neural networks and deep learning in the visual arts: A review. Neural Comput. Appl. 33, 121–157.

Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., Pantofaru, C.R., 2021. A step toward more inclusive people annotations for fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES.

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M., Lepri, B., 2017a. What your facebook profile picture reveals about your personality. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 460–468.

Segalin, C., Cheng, D.S., Cristani, M., 2017b. Social profiling through image understanding: Personality inference using convolutional neural networks. Comput. Vis. Image Underst. 156, 34–50.

Seguin, B., Striolo, C., diLenardo, I., Kaplan, F., 2016. Visual link retrieval in a database of paintings. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and (2016) 15-16, Proceedings, Part I 14. Springer, pp. 753–767.

Serengil, S.I., Ozpinar, A., 2021. Hyperextended lightface: A facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies. ICEET, IEEE, pp. 1–4.

Stork, D.G., Bourached, A., Cann, G.H., Griffths, R.R., 2021. Computational identification of significant actors in paintings through symbols and attributes. Electron. Imaging 33, 1–8.

Strano, M.M., 2008. User descriptions and interpretations of self-presentation through facebook profile images. Cyberpsychol. J. Psychosoc. Res. Cyberspace 2 (5).

Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M., 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nat. Mach. Intell. 3, 42–50.

Tores, J., Sassatelli, L., Wu, H.Y., Bergman, C., Andolfi, L., Ecrement, V., Precioso, F., Devars, T., Guaresi, M., Julliard, V., et al., 2024. Visual objectification in films: Towards a new ai task for video interpretation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10864–10874.

Ververas, E., Gkagkos, P., Deng, J., Doukas, M.Christos., Guo, J., Zafeiriou, S., 2022. 3DGazeNet: Generalizing gaze estimation with weak-supervision from synthetic views. arXiv e-prints, arXiv–2212.

Vilnai-Yavetz, I., Tifferet, S., 2015. A picture is worth a thousand words: Segmenting consumers by facebook profile images. J. Interact. Market. 32, 53–69.

Wevers, M., Smits, T., 2020. The visual digital turn: Using neural networks to study historical images. Digital Scholarship Humanities 35, 194–207.

Wijntjes, M., 2021. Shadows, highlights and faces: the contribution of a 'human in the loop'to digital art history. Art Percept. 9, 66–89.

Xing, L., Tian, Z., Huang, W., Scott, M.R., 2019. Convolutional character networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 9126–9136.

Yang, J., Zhou, K., Li, Y., Liu, Z., 2024. Generalized out-of-distribution detection: A survey. Int. J. Comput. Vis. 1–28.

Yao, L., Suryanarayan, P., Qiao, M., Wang, J.Z., Li, J., 2012. OSCAR: On-site composition and aesthetics feedback through exemplars for photographers. Int. J. Comput. Vis. 96, 353–383.

Ye, K., Kovashka, A., 2018. ADVISE: Symbolism and external knowledge for decoding advertisements. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 837–855.

Yi, R., Tian, H., Gu, Z., Lai, Y.K., Rosin, P.L., 2023. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22388–22397.

Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F., 2022. General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18697–18709.

Zhou, X., Koltun, V., Krähenbühl, P., 2022. Simple multi-dataset detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7571–7580.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al., 2023. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. arXiv preprint arXiv:2302.09419.