International Journal of Digital Humanities (2023) 5:451–490
https://doi.org/10.1007/s42803-023-00077-8

ORIGINAL PAPER

Hypericons for interpretability: decoding abstract concepts in visual data

Delfina Sol Martinez Pandiani[1] · Nicolas Lazzari[2] · Marieke van Erp[3] · Valentina Presutti[2]

Received: 15 March 2023 / Accepted: 10 November 2023 / Published online: 4 December 2023

Abstract

In an era of information abundance and visual saturation, the need for resources to organise and access the vast expanse of visual data is paramount. Abstract concepts-such as *comfort*, *power*, or *freedom*-emerge as potent instruments to index and manage visual data, particularly in contexts like Cultural Heritage (CH). However, the variance and disparity in the visual signals that evoke a single abstract concept challenge conventional approaches to automatic visual management rooted in the Computer Vision (CV) field. This paper critically engages with the prevalent trend of automating high-level visual reasoning while placing exclusive reliance on visual signals, prominently featuring Convolutional Neural Networks (CNNs). We delve into this trend, scrutinising the knowledge sought by CNNs and the knowledge they ultimately encapsulate. In this endeavour, we accomplish three main objectives: (1) introduction of ARTstract, an extensive dataset encompassing cultural images that evoke specific abstract concepts; (2) presentation of baseline model performances on ARTstract to elucidate the intricate nuances of image classification based on abstract concepts; and, critically, (3) utilization of ARTstract as a case study to explore both traditional and non-traditional avenues of visual interpretability, a trajectory inspired by Offert and Bell (2021). To more comprehensively understand how CNNs assimilate and reflect cultural meanings, and to discern the echoes reverberating within these visions, we unveil SD-AM, a novel approach to explainability that collapses visuals into hypericon images through a fusion of feature visualization techniques and Stable Diffusion denoising. Overall, this study critically addresses abstract concept image classification's challenges within the CNN paradigm. By embracing innovative methodologies and providing comprehensive analyses of explainability techniques, we make a substantial contribution to the broader discourse surrounding automatic high-level visual understanding, its interpretability, and the ensuing implications for comprehending culture within the digital era. Through our exploration, we illuminate the multifaceted

✉ Delfina Sol Martinez Pandiani
delfinasol.martinez2@unibo.it

Extended author information available on the last page of the article

trends, complexities, and opportunities that underlie the fusion of high-level visual reasoning and computer vision.

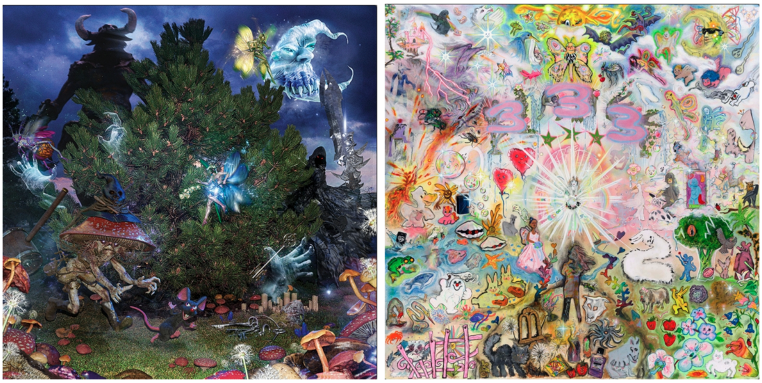
Keywords Computer vision · Explainability · Hermeneutics · Stable diffusion · Cultural heritage · Computational visual studies

1 Introduction

The proliferation of images in modern mass media has reached unprecedented levels, with social media platforms alone hosting billions of images every day, and Cultural Heritage institutions undergoing an unprecedented digitalisation of their image collections (Bevan, 2015). This phenomenon characterises the post-modern era, where users are overwhelmed by an abundance of information that is not curated (Jansson & Hracs, 2018), and increasingly latch to movements, such as *hyperpop* (Fig. 1), which strive to find some meaning amidst the disarray, by compiling vast fields of disparate meanings until they reach some semblance of accord (Vassar, 2020).

The conventional adage “knowledge is power” is slowly yielding to the prominence of curation in a contemporary shift from the information age to an era defined by thoughtful selection and arrangement. In this transition, the ability to sift through and adeptly manage copious data emerges as the genuine wellspring of influence (Rosenbaum, 2011; Cohen & Mihailidis, 2013; Jansson & Hracs, 2018). Unsurprisingly, automatic tools to navigate and comprehend vast troves of visual data have become paramount. Computer vision techniques, powered by Convolutional Neural Networks (CNNs), have emerged as indispensable tools in this endeavor to classify and categorize large collections of image data (LeCun et al., 2015), including for cultural images such as advertisements (Ye et al., 2019) and artworks (Cetinic & She, 2022). Simultaneously, there is a pursuit of automating increasingly intricate high-level visual reasoning tasks that go beyond concrete visuals–such as the prediction of personality traits (Segalin et al., 2017), political affiliation (Joo et al., 2014), and even facial beauty (Gray et al., 2010) from visual signals alone. These tasks, which for humans are deeply intertwined with cultural contexts and biases, have redefined the expectations placed upon computer vision models, the depth of knowledge they strive to acquire, and the crucial need for interpretability (Martinez Pandiani & Presutti, 2023).

In this context, abstract concepts (ACs) such as *comfort*,*freedom*, or *danger* become important tools for advancing the next generation of automated visual organization and curation. These concepts, which underlie the expression of emotions, opinions, and ideas through language (Kousta et al., 2011), hold significant influence in categorizing and managing visual data. This is because visual forms, such as paintings and photographs, function as vehicles of concepts, not solely by establishing links to depicted objects through visual attributes, but also through Roland Barthes’ notion of an image’s *connotation*-a secondary layer of meaning rooted in culturally encoded elements (Barthes, 1980). This is particularly evident in the domain of Cultural Heritage (CH), where controlled thesauri and classification systems frequently incorporate abstract concepts to categorize visual materials (Rafferty & Hidderley, 2017). Shared



(a) Mikey Joyce 2020[a] (b) Claire Barrow 2020[a]

Fig. 1 Hyperpop synthesizes divergent meanings, reflecting an era of information abundance and visual saturation

vocabularies and ontologies such as Iconclass,[1] Library of Congress,[2] Getty Vocabularies,[3] and Art and Architecture Thesaurus[4] offer pre-established ACs for association with visual content.

The potency of ACs in managing visual data lies in their capacity to bridge the gap between visual forms and cultural meanings, by activating heterogeneous scenes and situations (Borghi & Binkofski, 2014). However, this power also engenders the challenge of reconciling the disparities in the visual representation of abstract concepts. For instance, a single abstract concept can be evoked by widely diverse visual data, as in Fig. 2. The diversity and divergence present in visual signals evoking individual abstract concepts present a formidable challenge to conventional methodologies rooted in the Computer Vision (CV) domain. Although CNNs offer immense promise, they are intrinsically optimized for tasks characterized by high intra-class similarity (Benz et al., 2020; Shirali & Hardt, 2023)—qualities that conflict with the inherent heterogeneity and culturally nuanced nature of abstract concepts. The task of detecting abstract concepts within visual data thus can be seen as a complex “wicked problem”, lacking clear-cut solutions and molded by multifaceted cultural intricacies (Rittel, 1967). The increasing utilization of CNNs for wicked problem tasks like abstract concept detection raises questions about the knowledge assimilated by these models. Thus, the explainability of these models becomes crucial in understanding how they handle complex socio-cultural visual reasoning tasks.

Explainability can be seen as a response to the perceived “explanatory deficit” in technical disciplines (Berry, 2021). It stems from the challenges posed by opaque models and the recognition of problematic biases that can lead to inequities. The emerging subfield of eXplainable AI (XAI) advocates for interpretability as a means to address these issues (Mittelstadt et al., 2016). In this context, the ability to explain predic-

[1] https://iconclass.org/

[2] https://www.loc.gov/

[3] https://www.getty.edu/research/tools/vocabularies/

[4] https://www.getty.edu/research/tools/vocabularies/aat/index.html

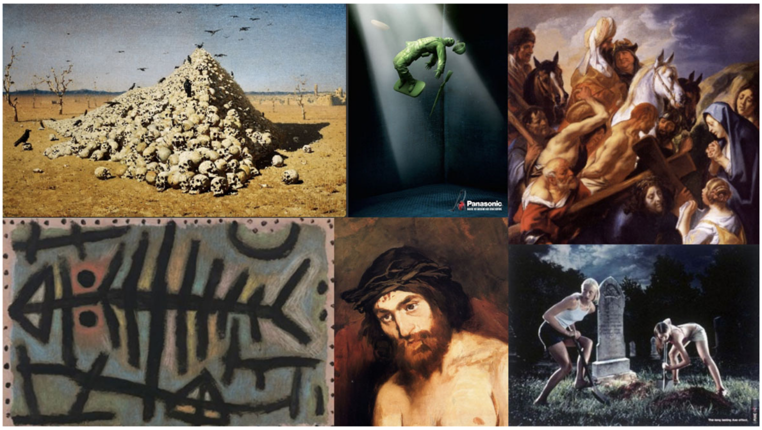


Fig. 2 Visual disparity in images that evoke the abstract concept of "death", collected in our ARTstract dataset. Clockwise from top left: *The Apotheosis of War* (1871) by Vasily Vereshchagin, Tretyakov Gallery, Moscow, Russia, public domain; *Panasonic: Where no vacuum has gone before* (2014) advertisement by Y&R; *Christ Carrying the Cross* (1660) by Jacob Jordaens, Rijksmuseum, Amsterdam, Netherlands, public domain; *The Axe Effect* (2003) advertisement by Lowe Bull Calvert Pace; *The head of Christ* (1864) by Edouard Manet, public domain; *Mess of fish* (1940) by Paul Klee, public domain. Images source for the Ads Dataset (Hussain et al., 2017) and the Artemis Dataset (Achlioptas et al., 2021)

tions is crucial and informative of the heuristic process itself Offert and Bell (2021). We argue that for the socio-cultural-cognitive task of abstract concept-based image classification (AC image classification)–which is based on subjective, cultural, and interoceptive processes–the perils of model reuse without explainability are high, as it can potentially echo harmful stereotypes or visions of the world based on prejudice, racism, and other biased worldviews.

In the context of machine vision, we are especially interested in the explainability techniques of class activation mapping (CAM) and activation maximization (AM), also known as feature visualization (FV). CAM identifies the salient regions of an image that are considered important by the model in a classification task. It has been used, for instance, to identify where CNN models localise symbols in iconography classification (Vago et al., 2021). FV (AM), on the other hand, involves the generation of images that visualize what a neuron in a CNN has learned (Erhan et al., 2009). It has been employed to visualize neurons of models trained on natural images as well as on artistic images (see Fig. 3). Offert and Bell (2021) contend that they can be viewed as technical metapictures (Mitchell, 1995) functioning as *hypericons*: indirect "illustrations," or "visualizations" in the literal sense of forcibly and subjectively summarizing and making-visual the non-visual, which can serve as summary images "a theory of knowledge" (Mitchell, 1995 p. 9) becoming crystallized in a machine learning model. Following this idea, Offert (2019) suggest a non-traditional approach to visual explainability that involves examining a large set of feature visualization images when performing dataset analysis, so that the hermeneutic work extends back into the technical system itself, operating on both the original dataset and the feature

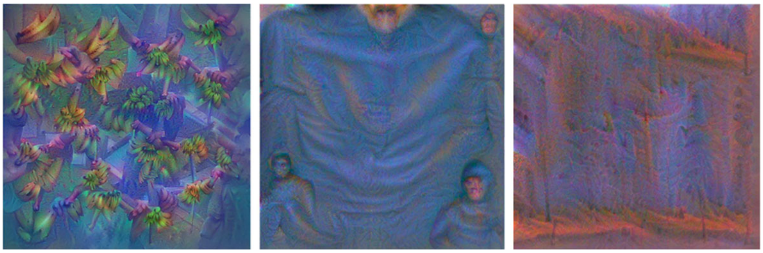


Fig. 3 Examples of regularized feature visualizations (FV) showing what specific neurons have learned. Left: FV for the *banana* class in an Inception-based CNN trained on ILSVRC2012; image credit: adapted from Offert and Bell (2021). Center and right: FVs for *portrait* (center) and *landscape* (right) classes in an Inception-based CNN finetuned on an art historical dataset to distinguish between portrait and landscape classes, trained on ImageNet and finetuned for 10 epochs on an art historical dataset, image credit: adapted from Offert (2019)

visualization dataset. Through the use of such methods, the technical system becomes an integral part of the interpretive process rather than an opaque tool.

This paper takes a critical stance towards the prevalent trend of high-level visual reasoning, driven by Convolutional Neural Networks (CNNs) reliant on visual signals. The investigation primarily targets abstract concept detection and the explainability of CNNs within this context, guided by three pivotal research points:

- **Dependability on the visual signal**: We probe the reliability of the decontextualized visual signal, as used by CNNs, for AC image classification, to assess the extent to which low-level features can be harnessed for this complex high-level visual task.
- **Insights from explainability techniques:** We apply traditional explainability techniques, particularly Grad-CAM++, to illuminate the nuances of CNN behaviour and decision boundaries in AC image classification.
- **Hypericons as a non-traditional explainability approach**: We present a case study of the “hypericons via feature visualizations” approach to visual interpretability, inspired by and extending (Offert, 2019; Offert & Bell, 2021), i.e. operating on both the original dataset and a derived feature visualization dataset of ‘hypericons’.

The technical contribution of this research can be summarized as follows:

1. formulation and development of the ARTstract dataset,[5] a dataset of art images and their evocation of certain abstract concepts;
2. baseline models and performances on ARTstract for the task of AC image classification;
3. introduction of the novel stable diffusion-denoised activation maximization (SD-AM) approach to hypericon creation. SD-AM facilitates the creation of human-understandable “hypericons,” unlocking the potential to uncover prototype images encapsulating associations with abstract concepts.

The remainder of this paper is organised as follows. In Section 2, we review related work, including computer vision work in relation to image classification, digi-

[5] https://github.com/delfimpandiani/ARTstract_Seeing_abstract_concepts

tal humanities, and explainability, as well as related work on explainability and digital humanities. We also discuss related work on abstract concepts and computer vision. In Section 3, we introduce the ARTstract dataset, including the original data sources, abstract concept selection, image selection, data processing, and dataset integration. In Section 4, we present abstract concept-based image classification baselines, including the experimental setup and the results. In Section 5, we discuss our explainability experiments, including the results from class feature visualizations and their denoising, as well as GradCAM++ feature visualizations. In Section 6, we provide a comprehensive discussion of the results, with a focus on contributions, lessons, and future directions in Section 6. We conclude in Section 8.

2 Related work

To contextualize our study, we examine how the domains of computer vision, explainable AI, and digital humanities intersect and contribute to our understanding of AC image classification and its challenges.

2.1 Computer vision for image classification

The introduction of ILSVRC (Russakovsky et al., 2015), a large-scale image classification challenge on ImageNet (Deng et al., 2009), marked the introduction of ever-improving image classification models. Convolutional Neural Networks (CNNs) represent one of the most widely-used methods in image classification tasks and are the backbone of modern state-of-the-art methods (Chen et al., 2021). A CNN is composed of several convolutional layers. A convolutional layer learns how to filter an image by learning the filter's kernel. By computing the convolution with the learned kernel over the whole image, the network extracts relevant features for the classification task. CNNs can be classified into three main classes: classical CNNs, inception CNNs and residual CNNs (Chen et al., 2021). Classical CNNs, such as VGG (Simonyan & Zisserman, 2015), make straightforward use of convolutional layers. Better performances are achieved using deep network models (i.e. networks with a large number of convolutional layers). The use of increasingly deep networks, however, has been shown to increase performances only to a certain extent (Chen et al., 2021). To overcome this limitation inception-based methods, such as InceptionNetV3 (Szegedy et al., 2016), and residual CNNs, such as ResNet (He et al., 2016), have been proposed. Recently, following the success of the Transformer architecture (Vaswani et al., 2017; Lin et al., 2022), transformer-based classification models, such as ViT (Zhai et al., 2022), have been introduced with promising results.

2.2 Computer vision and digital humanities

Whilst much computer vision (CV) work is focused on analysing digital-born, realistic photos, several research groups are investigating computer vision for (art-) historical datasets (cf. Stork, 2009; Rodríguez-Ortega, 2020). As art, cultural and historical images are not always photo-realistic, transfer learning is often applied to deal with differences in style for example to handle paintings (Zinnen et al., 2023) or photos in historical newspapers (Wevers & Smits, 2020). Furthermore, the types of objects that are relevant for naturalistic photos (e.g. aeroplanes, cars, etc.) are not always relevant to tasks in the humanities domain, therefore new datasets and models are trained for this domain. To process medieval manuscripts, one for example needs to detect images placed in running text, that present stylised depictions of objects (Bekkouch et al., 2021). Different research questions demand different objects to be detected, such as 'smelly' objects (Zinnen et al., 2022), zoological species (Stork et al., 2021), railway accidents (Smits, 2023) or musical instruments (Sabatelli et al., 2021). In general, the perspective humanists bring to CV is that the CV methods are a tool to answer a research question rather than the objective of the research itself (Wevers & Smits, 2019; van Noord, 2022). Lately, as the humanities domain has a longstanding tradition of source criticism (Régimbeau, 2014), which more recently was expanded to tool criticism (Koolen et al., 2019), there has been a fair amount of attention for biases in datasets and algorithms (Wevers, 2019; Offert & Bell, 2021; Smits & Wevers, 2022).

2.3 Computer vision and explainability

In this section, we succinctly describe two of the most commonly used techniques for *post-hoc* explainability of CNN-based computer vision models (Vilone & Longo, 2020; Ibrahim & Shafiq, 2023), class activation mapping (CAM) and activation maximisation (AM).

2.3.1 Class activation mapping (CAM)

An approach to the visual explanation of image classification models is CAM (Zhou et al., 2016), a method initially proposed to investigate how CNN models trained on classification tasks are able to generalise on localisation tasks. On the XAI landscape, CAM methods are used to highlight the regions of an image that are important in the classification process of a model (Ibrahim & Shafiq, 2023). Given the output of the last convolutional layer of a CNN that classifies an image, the one that displays the higher spatial resolution when compared to other layers (Zhou et al., 2016), a visual explanation map can be computed by enhancing the response of highly-activated neurons. Different enhancements methods have been proposed: GradCAM (Selvaraju et al., 2020) and GradCAM++ (Chattopadhyay et al., 2018) use the layer's gradient to compute coefficients; XGrad-CAM (Fu et al., 2020) uses of axioms to avoid the use of heuristic methods; LIFT-CAM (Jung & Oh, 2021) proposes an analytical solution to the problem. CAM methods (Vilone & Longo, 2020; Ibrahim & Shafiq, 2023) represent one of the main methods used to obtain a visual explanation of image classification models.

2.3.2 Activation maximisation (AM)

A different approach to the visual explanation of image classification models is the Activation Maximisation (AM) method (Erhan et al., 2009; Vilone & Longo, 2020), where the activation of the classification neurons is exploited to synthesise “prototypical” images of a class. AM is formulated as an optimisation problem, where an image is obtained by directly maximising the activation of one or more classification neurons (Nguyen et al., 2019). The optimisation procedure can be expressed in terms of gradient ascent when the gradient is accessible (Erhan et al., 2009; Nguyen et al., 2016, 2019). The resulting image is often hard to interpret from a human point of view. Different regularisation techniques have been proposed to address this issue, such as L_α norm (Simonyan et al., 2013) to smooth pixel intensities, Total Variation (TV) (Mahendran & Vedaldi, 2016) to encourage smoothness of the image (Nguyen et al., 2019) and gradient enhancing techniques (Bykov et al., 2022). A different approach to obtain more realistic images and human-understandable images is to synthesise the image using a generator network, as done in DGN-AM [6] (Nguyen et al., 2016) by using a Generative Adversarial Network (GAN). This biases the image towards more realistic-looking images that are easier to interpret from a human’s point of view. Recently, similar approaches have been used to guide the generation process of Diffusion Models [7] (Dhariwal & Nichol, 2021). Such methods have shown promising results in providing *post-hoc* explanations of classification models (Jeanneret et al., 2022) but have not been applied to the AM process.

2.4 Explainability and digital humanities

Due to the more reflective nature of humanities research, explainability and trustworthiness are core concerns for many humanities scholars (Berry, 2022). In a way, computer science has borrowed concepts that contribute to explainability from the humanities domain, such as *provenance*. There is now a vibrant computational provenance community in computer science (Deutch et al., 2022), but this concept originates from (art-)history (Moreau et al., 2008). Humanities researchers are questioning the explainability of digital methods and the impacts these can have on conclusions drawn, in for example visualisations (Boyd Davis et al., 2021), and hermeneutics (Van Zundert, 2015). Some humanists take it upon themselves to dive into the details of digital methods in an attempt to understand the impacts of different parameter settings (Evert et al., 2017; Wevers & Smits, 2019; van Lange, 2022), but this is not possible for everyone. Methods such as tool criticism (Koolen et al., 2019) can be instrumental in furthering explainability for digital humanities. An additional view on improving explainability in digital humanities is to provide as much of the data and code as possible, for example through new publication methods such as Jupyter notebook-based articles in the Journal of Digital History.[8]

[6] https://github.com/Evolving-AI-Lab/synthesizing

[7] https://github.com/openai/guided-diffusion

[8] https://journalofdigitalhistory.org/en

2.5 Abstract concepts and computer vision

Detecting Abstract Concepts (ACs) from images is a challenging task for computer vision (Hussain et al., 2017) due to subjectivity, class imbalances, and entropy. Specifically, ground-truth generation is highly inconsistent, caused by personal and situational factors such as cultural background, personality, and social context (Zhao et al., 2018). Almost all trainable classes in popular datasets for image classification, such as ImageNet[9] and Google Open Images,[10] refer to concrete classes (Abgaz et al., 2021; Ahres & Volk, 2016; Brigato et al., 2022). Moreover, abstract words tend to have higher dispersion ratings due to the wide variety of images returned from a query (Kiela & Bottou, 2014; Lazaridou et al., 2015). Despite these issues, a few works define their task as grappling with abstract concept detection from images (Abgaz et al., 2021; Ye et al., 2019; Kalanat & Kovashka, 2022). For example, Kalanat and Kovashka (2022) builds on previous work (Ye et al., 2019) focusing on classifying advertisement images based on symbol clusters.

Related work with similar goals in diverse subfields of computer vision includes visual sentiment analysis (Ortis et al., 2020) in artistic (Zhao et al., 2021), meme (Sharma et al., 2020) and natural images, including specifically those of groups of people (Veltmeijer et al., 2021). Social signal processing has attempted to detect non-concrete social aspects of images, such as persuasive intent, deception, social relationship type, intimacy, and non-concrete aspects of groups, among others (Joo et al., 2014; Kantharaju et al., 2020; Solera et al., 2017; Zhang et al., 2018; Varghese & Thampi, 2018; Vanneste et al., 2021).

To the best of our knowledge, the only publicly available dataset of annotated images which has a focus on non-concrete nominal concepts is BabelPic (Calabrese et al., 2020), which presents a methodology to automatically produce a silver dataset that extends the BabelPic coverage to all BabelNet synsets. Popular datasets such as ImageNet offer little to no coverage of non-concrete concepts, and JFT, an internal dataset at Google, includes non-concrete classes but is not publicly released (Deng et al., 2009; Sun et al., 2017). Datasets containing at least some images annotated with non-concrete labels can be categorized by image type. For natural images, NUS-WIDE, Google's Open Images dataset, MultiSense, VerSe, and Persuasive Portraits of Politicians include partial coverage of AC labels (Chua et al., 2009; Kuznetsova et al., 2020; Gella et al., 2019, 2016; Joo et al., 2014). However, most of the abstract concepts are not trainable classes due to their low instances. Finding datasets of art and cultural images is even more challenging. Crucial ones include WikiArt Emotions Dataset, which contains annotations for emotions evoked in the observer, and which overlaps with ARTemis, which contains emotion attributions and explanations on artworks also from WikiArt, including associations with abstract concepts (Mohammad & Kiritchenko, 2018; Achlioptas et al., 2021). The Ads Dataset, introduced by Hussain et al. (2017), links symbolic (abstract) concepts to specific regions within images, making it unique in the field. ArtPedia (Stefanini et al., 2019) provides a dataset of 2930 images of paintings with both visual and contextual descriptions, some of which

[9] https://www.image-net.org/

[10] https://storage.googleapis.com/openimages/web/index.html

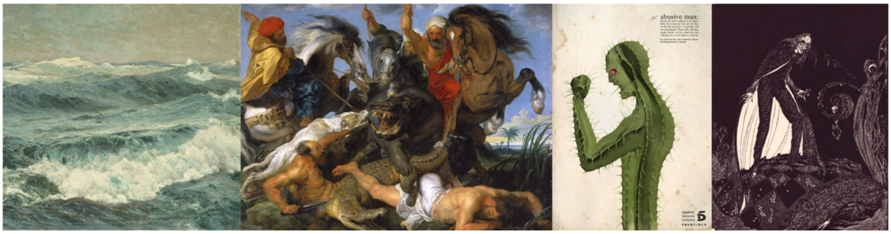


Fig. 4 Four instances of visual cultural images tagged with the abstract concept *danger* in ARTstract. This example clearly shows that the abstract concept labels are semantically diffused and associated with visually variant images. From left to right: *The Roaring Forties* (1908) by Frederick Judd Waugh, Metropolitan Museum of Art, public domain; *The Hippopotamus and Crocodile Hunt* (1615) by Peter Paul Rubens, Alte Pinakothek in Munich, Germany, public domain; Untitled advertisement by Telecinco against domestic violence; *Tales of Mystery and Imagination by Edgar Allan Poe* (1923) by Harry Clarke, Metropolitan Museum of Art, public domain. Images sourced from the Ads Dataset (Hussain et al., 2017), the Artemis dataset (Achlioptas et al., 2021), and ArtPedia (Stefanini et al., 2019)

contain abstract concepts. Finally, the Tate Gallery collection metadata of 70,000 artworks are tagged with Tate's subject taxonomy, which includes both concrete and social concepts as subject tags.

3 ARTstract dataset

In alignment with our exploration of the intricate challenges posed by abstract concepts within the realm of computer vision and computational visual studies, we introduce the ARTstract dataset, a resource that addresses the gaps and complexities discussed thus far (c.f. Fig. 4). This dataset presents a tool for researchers seeking to delve into the intersection of visual data and abstract ideas. Comprising an array of high-resolution images encompassing cultural artworks and advertisements, each associated with abstract concepts, ARTstract emerges as a significant asset to advance our understanding of the interplay between visual content and conceptual meaning. A more complex, detailed, and knowledge-graph-based version of the dataset is under construction.[11]

ARTstract was curated by combining sections from four distinct image datasets, namely ArtPedia, ARTEMIS, the Ads Dataset, and the Tate Collection (see Section 3.1 for a comprehensive overview). The inclusion criteria for images were twofold: relevance to abstract concepts and high resolution. By adopting this approach, ARTstract encapsulates a diverse range of visual materials that have inherently interwoven abstract ideas within their imagery.

3.1 Original data sources

Artemis Achlioptas et al. (2021) is a large-scale dataset providing data about the interplay between visual content, its emotional effect, and its language explanations.

[11] https://github.com/delfimpandiani/ARTstract-KG

Artemis focuses on the affective experience triggered by visual artworks and asks annotators to indicate the dominant emotion they feel for a given image, as well as to provide a grounded verbal explanation for their emotion choice. The dataset contains 455K emotion attributions and explanations from humans on 80K artworks from WikiArt. It provides a rich set of signals for both the objective content and the affective impact of an image, creating associations with abstract concepts, such as "freedom" or "love," or references that go beyond what is directly visible, including visual similes and metaphors or subjective references to personal experiences.

Artpedia Stefanini et al. (2019) provides a dataset of paintings with both visual and contextual descriptions. The dataset contains over 2,930 images, and the manual annotation of each sentence as either visual or contextual allows for a comprehensive analysis of the visual and semantic content of the dataset. Additionally, Artpedia is the only dataset to contain both types of artistic sentences, making it a valuable resource for developing visual-semantic models capable of jointly discriminating between visual and contextual sentences of the same painting. Some of the dataset's visual sentences include abstract concept terms, making it a valuable resource for researchers interested in abstract concepts in the cultural heritage field.

The Tate Gallery houses the United Kingdom's national collection of British art, as well as international modern and contemporary art. Their collection metadata of 70,000 artworks, available as a Github repository,[12] includes complete records of most artists and artworks in the collection, along with image and thumbnail URLs. The Tate's subject taxonomy for labelling their artworks is a unique feature of the dataset, as it includes both concrete and social concepts as subject tags. The taxonomy was expert-led, developed alongside the digitization of the Tate's collection, and organized in a hierarchical structure.

The Ads Dataset (ADVISE) Hussain et al. (2017); Ye and Kovashka (2018) includes over 64,000 image ads covering a diverse range of subjects. Each image ad is tagged with its topic, the sentiment it attempts to inspire in the viewer, and the strategy it uses to convey its message. The dataset also includes "symbols" that the ads use, a common technique used in advertising to convey meaning and emotions to the viewer, such as the concept of "peace" symbolically represented by a dove. The dataset includes 13,938 ad images with 221 abstract concept symbols, each with corresponding bounding boxes, and then clustered into 53 symbol clusters. The most common symbolic abstract concepts are "danger," "fun," "nature," "beauty," "death," "sex," "health," and "adventure."

3.2 Abstract concept selection and definition

The ARTstract dataset uses evoked clusters as a way to label the abstract concepts present in each image. Evoked clusters are groups of abstract concepts that often co-occur together in a given context. The idea of clustering abstract concepts or symbols

[12] https://github.com/tategallery/collection

for their visual evocation was introduced alongside the Ads Dataset (Hussain et al., 2017), and has been used in the little existing research on AC image classification in the context of computer vision, such as Ye and Kovashka (2018) and Kalanat and Kovashka (2022). While the cluster categories are not perfect nor objective, the only available performances for the task of AC image classification use these clusters, which is why we have decided to reuse them to create this dataset. The original clusters were created by analyzing the co-occurrence of abstract concepts in advertisement images. Certain choices that were made by the creators of these clusters are decidedly Western-oriented and may not be shared in all contexts. We advocate for future work to investigate more culturally-sensitive and/or diverse abstract concept datasets.

In selecting AC clusters for our study, we drew upon the insights from cognitive science research. Specifically, we cross-referenced the clusters from Hussain et al. (2017) with a list of abstract concepts from recent foundational work in abstract concepts in cognitive science (Harpaintner et al., 2018). By leveraging cognitive science research, we aimed to identify a diverse set of ACs that would facilitate our investigation into the relationship between abstract concepts and visual imagery. We then excluded the ACs that have been previously explored in CV literature, including emotions (such as *happiness*, and *love*) and *violence* (Ramzan et al., 2019). After a thorough selection process, we identified eight clusters of ACs, with cluster words defined by Hussain et al. (2017) that we deemed appropriate for our research, namely:

- comfort: *comfort, cozy, soft, softness*
- danger: *danger, peril, risk*
- death: *death, lethal, suicide, funeral*
- excitement: *excitement, flavors*
- fitness: *exercise, fitness, running*
- freedom: *america, freedom, liberty*
- power: *force, power, powerful*
- safety: *safety, security*

3.3 Image mining and processing

We then searched through various art and cultural image datasets to match the clusters with relevant images. To do so, we followed different steps for each of the four original data sources. Succintly, in the Tate and ADVISE datasets, individual images have individual word labels assigned to them, including terms that refer to abstract concepts. As such, for each AC cluster, we identified the images that had been tagged with at least one of the cluster words, and also kept track of how many times one of that clusters' words had been used (i.e., the evocation strength). For ARTEMIS and ArtPedia, we mined the "utterances" and "visual sentences", respectively, to identify if they contained any of the cluster evokers. If they did, we considered that an evocation, and, as for the other datasets, we kept track of the number of evocations to then calculate the evocation strength. The complete and documented steps can be found in the Github repository.

3.4 Dataset integration

Each image in the ARTstract dataset is assigned one or more evoked clusters based on the abstract concepts present in the image. For each assignment, words associated with the image, evocation strength, and evocation evidence are tracked alongside the identified cluster.

The dataset contains 16,166 images, each with one out of eight abstract concept clusters as tags. The images are in JPG format and have a resolution of 512x512 pixels. Each image is labelled with the evoked AC cluster. Additionally, the dataset is accompanied by a knowledge base where each image has is connected to evocation information such as the context or evidence of evocation as well as the evocation strength.

The dataset suffers from a class imbalance that could hinder models' ability to generalise accurately (as seen in Fig. 5). To address this, we extract a balanced subset of the dataset, which we call the *balanced ARTstract*. This involves randomly selecting a maximum of 1,000 images from classes with more than 1,000 instances while retaining all images for classes with fewer than 1,000 instances.

4 Performance experiments

In this section, we present our performance methodologies, experimental setup, and results. Recognizing the challenges ingrained in the task of AC image classification, and hypothesizing low performances given the inherent problem of high intra-class variance for abstract concepts, we still decided to train a dedicated model for this task. This choice stems from the desire to investigate the effectiveness of CNNs in AC classification, the need to explore potentially relevant low-level perceptual features and investigate the models' ability to capture prototype images for these concepts. We further discuss our approach's rationale and implications in Section 6.

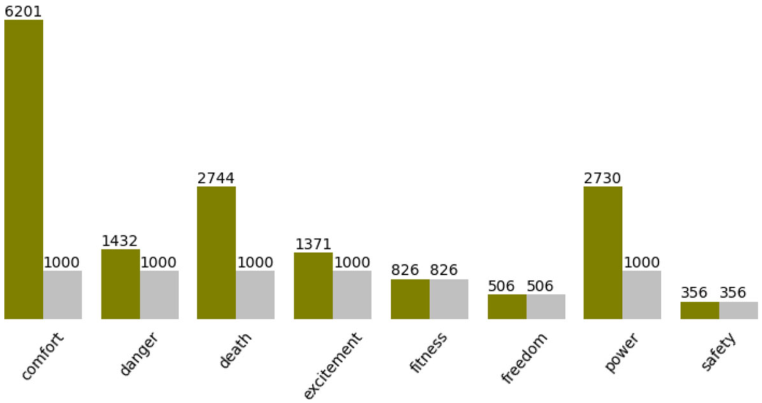


Fig. 5 ARTstract dataset distribution: the unbalanced version (green) contains a total of 16166 images, while the balanced one (grey) contains 6688

4.1 Methodology

We want to assess whether convolutional neural networks (CNNs) are suitable models to achieve adequate performance in the task of classifying images according to the abstract concepts they evoke. We refer to this task as *abstract concept-based image classification* (AC image classification).

We experiment on the ARTstract dataset (introduced in Section 3) by re-using models trained on ILSVRC (ImageNet), since pre-training a CNN on large-scale data has been shown to result in more accurate results (Kornblith et al., 2019). Indeed, the inductive bias injected into the convolutional layers establishes a set of filters that are effective at detecting prominent features from images. Even though the classification task of models trained on ImageNet is different to the one of ARTstract, the images can be considered perceptually similar. We argue that the extraction of such prominent features (e.g. detecting edges) and high-level features (e.g. identifying objects) is shared between both datasets. For this reason, we only fine-tune the classification layer of the model while keeping the rest of the layers frozen. This minimises the impact of the limited amount of data available, as it has been shown that training CNN from scratch on low-volume data results in degraded performances (Sharma & Mehra, 2018).

For the fine-tuning procedure, we follow ILSVRC multi-class formulation (Russakovsky et al., 2015). Formally, we estimate the probability $p(c \mid \hat{x}, \Theta)$ where c is among the classes (C) presented in Section 3, $\hat{x}$ the input image and Θ is the neural network used to parametrise the probability distribution, in our case a CNN. Differently than ILSVRC, where the top-k predicted classes are evaluated, we compute accuracy (A), precision (P), recall (R) and F1 defined as

$$A = \frac{1}{\mid C \mid} \sum_{c \in C} \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \qquad P = \frac{1}{\mid C \mid} \sum_{c \in C} \frac{TP_c}{TP_c + FP_c}$$

$$R = \frac{1}{\mid C \mid} \sum_{c \in C} \frac{TP_c}{TP_c + FN_c} \qquad F1 = \frac{1}{\mid C \mid} \sum_{c \in C} \frac{P_c \cdot R_c}{P_c + R_c}$$

where TP_c, TN_c are, respectively, the correct prediction of an image in and not in a class and FP_c, FN_c are, respectively, the wrong prediction of an image in and not in a class. We rely on different measures since there is a consistent difference between the number of classes on ImageNet (1000) and the classes on ARTstract (8). Evaluating the top-k results (e.g. top-5 accuracy) would result in over-optimistic results.

We experiment with two CNNs that serve as baseline models, respectively VGG-16 (Simonyan & Zisserman, 2015) and ResNet-50 (He et al., 2016). The main difference between the two architectures is the number of their convolutional layers: 16 for VGG-16 and 50 for ResNet-50. We want to assess the performance of these models in the AC image classification task as well as to investigate whether bigger models (i.e., models that employ a larger number of convolutions) obtain better performances, as already observed in different image classification tasks (Chen et al., 2021).

Table 1 Hyperparameters used to train the classification baselines

Finetuned Model	Epochs	Batch Size	Learning Rate
ResNet-50	100	32	0.001
VGG-16	100	32	0.001
VGG-16	1000	32	0.001

4.2 Experimental setup

We train each model using an RTX3090 on an Intel i9 CPU with 8 cores and equipped with 128 GB of RAM. We manually adjust and experiment with different hyperparameters, summarised in Table 1, and train using the Adam optimizer (Kingma & Ba, 2015), on the whole dataset split into train, validation, and test sets using an 80:10:10 ratio.

To further reduce overfitting e employ standard data augmentation (Shorten & Khoshgoftaar, 2019) such as random horizontal flips, random color jitter, random rotation, and random crop. For each image in the training set, we resize it to 224x224 pixels, apply a random horizontal flip with a probability of 0.5, a random color jitter with a probability of 0.3, a random rotation with a probability of 0.3, and a random crop of size 20 with a probability of 0.3 and subtract ImageNet mean color. For the images in the validation and testing dataset, we only resize to 224x224 pixels and subtract ImageNet mean color. Given the described methodology, we are able to train on the whole ARTstract dataset, minimising the overfitting due to the skewed distribution of the dataset (c.f. Fig. 5).

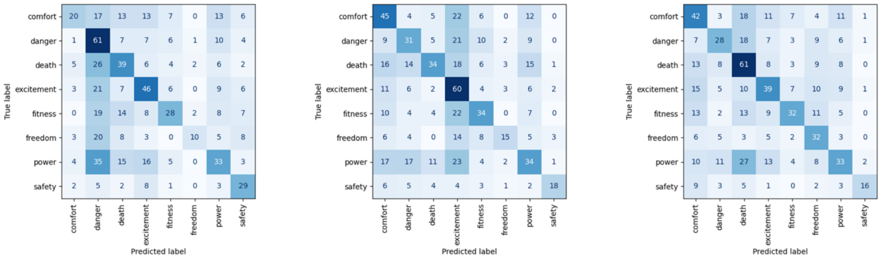
4.3 Results

We finetune the chosen baseline models as described above. In Table 2 we report the performance metrics of the tested models. VGG-16 finetuned on 100 epochs outperforms ResNet-50 in all the described measures. To further investigate the potential of VGG-16 we finetune the model for 1000 epochs. The resulting model outperforms all the other baseline models. Figure 6 depicts the confusion matrices for the models of Table 2. Interestingly, all models are prone to misclassification errors on the *power* class. While this aspect requires an in-depth analysis, to investigate whether it

Table 2 Comparison of model performances between a pretrained ResNet-50 finetuned for 100 epochs, a pretrained VGG-16 finetuned for 100 epochs, and a pretrained VGG-16 finetuned for 1000 epochs on the test set

Finetuned Model	Loss	Accuracy	Precision	Recall	F1
ResNet-50 - 100 epochs	1.732	0.392	0.451	0.392	0.419
VGG-16 - 100 epochs	1.626	0.402	**0.465**	0.402	0.431
VGG-16 - 1000 epochs	**1.605**	**0.426**	**0.465**	**0.426**	**0.445**

Best performance for each metric is highlighted in bold



(a) ResNet-50 trained for 100 epochs.

(b) VGG-16 trained for 100 epochs.

(c) VGG-16 trained for 1000 epochs.

Fig. 6 Confusion matrices for the baseline models pre-trained on ImageNet and finetuned on balanced ARTstract

is due to mislabeled images in the dataset, we argue that it be partially reconducted to our *implicit* definition of *power*, which might have little correlation with the bias introduced by ImageNet. The most problematic class is not consistent between the different models. ResNet-50 (Fig. 6a), VGG-16 finetuned for 100 epochs (Fig. 6b) and VGG-16 finetuned for 1000 epochs (Fig. 6c) overfit respectively on the classes *danger*, *excitement* and *death*. In all cases, this accounts for most of the misclassification errors. More extensive augmentation techniques, longer training sessions, and more sophisticated models might help in overcoming this issue, which we will explore in future works.

5 Explainability experiments

Inspired by Offert (2019); Offert and Bell (2021), we experiment with three different techniques to obtain explainable results from the baseline models presented in Section 4: CAM, AM, and SD-AM.

5.1 Methodology

For each explanation experiment, we investigate the classification process of the best model from Table 2 of Section 4, VGG-16 finetuned for 1000 epochs. We experiment with GradCAM++ (Chattopadhyay et al., 2018) for the CAM method and Stable Diffusion (Rombach et al., 2022) for SD-AM. Results are evaluated manually.

Class activation mapping

We are interested in investigating which parts of an image are mostly influencing the classification of our fine-tuned models. We use the Class Activation Mapping (CAM) method which, given $f_c^l(x)$ the output of the l-th layer of a CNN that classifies the image x with the class c, computes a visual explanation map as

$$CAM(f_c^l(x)) = ReLU(\sum_{i=0}^{N_l} \alpha_k \cdot f_c^l(x)_k)$$

where N_l is the number of channels of the l-th layer, $f_c^l(x)_k)$ is the output of the k-th channel of the l-th layer and α_k is an hyper-parameter of the model. The last convolutional layer is usually taken as l, since it has been shown to display a higher spatial resolution when compared to other layers (Zhou et al., 2016). GradCAM++ (Chattopadhyay et al., 2018) uses the layer's gradient to compute coefficients. Since a precise localisation is not the primary focus of our research, but we are instead interested in highlighting the approximate regions of interest for the model for a specific classification, we manually experiment with different methods and decide to rely on GradCAM++ as implemented in `pytorch-grad-cam`[13] (Gildenblat, 2023). We are interested in investigating which parts of an image are mostly influencing the classification of our fine-tuned models. Since a precise localisation is not the primary focus of our research, but we are instead interested in highlighting the approximate regions of interest for the model for a specific classification, we manually experiment with different methods and decide to rely on GradCAM++ as implemented in `pytorch-grad-cam`[14] (Gildenblat, 2023). We generate saliency maps for images of interest using our best model, VGG-16 finetuned for 1000 epochs and manually inspected the results (for example, Fig. 8 to obtain valuable insights into the classification criteria of the model.

Activation maximization

To investigate the perceptual topology of the model, we decided to generate Activation Maximization (AM) images for the neuron responsible for a specific class. AM can be formulated as an optimisation problem, with the objective function is defined as

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, a_c(x) \tag{1}$$

where $\hat{x}$ is an image that maximises the neuron a_c responsible for classifying as a class c. To generate the AM images, we rely on OmniXAI's[15] (Yang et al., 2022) feature visualization implementation; we experiment with combinations of different parameters, as described in Table 3.

Stable diffusion-activation maximization

Finally, inspired by the work of DGN-AM (Nguyen et al., 2016), where the authors obtain realistic-looking images using a GAN generator, and given the recent success

[13] https://github.com/jacobgil/pytorch-grad-cam

[14] https://github.com/jacobgil/pytorch-grad-cam

[15] https://github.com/salesforce/OmniXAI

Table 3 Activation Maximization parameters supported by OmniXAI (Yang et al., 2022) library

Parameter	Values
Iterations	300, **400**, 500
Learning rate	**0.1**, 0.01, 0.01
Regularizer	L_1, L_2, **TV**
Regularizer weight	0, −0.05, −0.5, **-2.5**
Fourier preconditioning	**yes**, no
Map uncorrelated colors to normal colors	**yes**, no

The best parameters after manual inspections are represented in bold

of diffusion models in the automatic image generation task (Dhariwal & Nichol, 2021; Rombach et al., 2022; Ramesh et al., 2022), we experiment on the same task by using Stable Diffusion (SD)[16] (Rombach et al., 2022), a diffusion-based image generator model, to synthesize realistic images from the ones obtained using the AM method (this method is hereto referred to as SD-AM). Informally, diffusion-based models progressively remove noise from an image using a neural network (Ho et al., 2020). The denoising procedure is generally guided by a textual prompt. We exploit this aspect by treating the image produced by the AM method as a noisy image and gradually removing noise from it using Stable Diffusion. We experiment in two different settings:

- Denoise the AM image without providing any textual prompt;
- Denoise the AM image by using a textual prompt as well, composed of the class label of the image (i.e. *comfort*, *danger*, etc.).

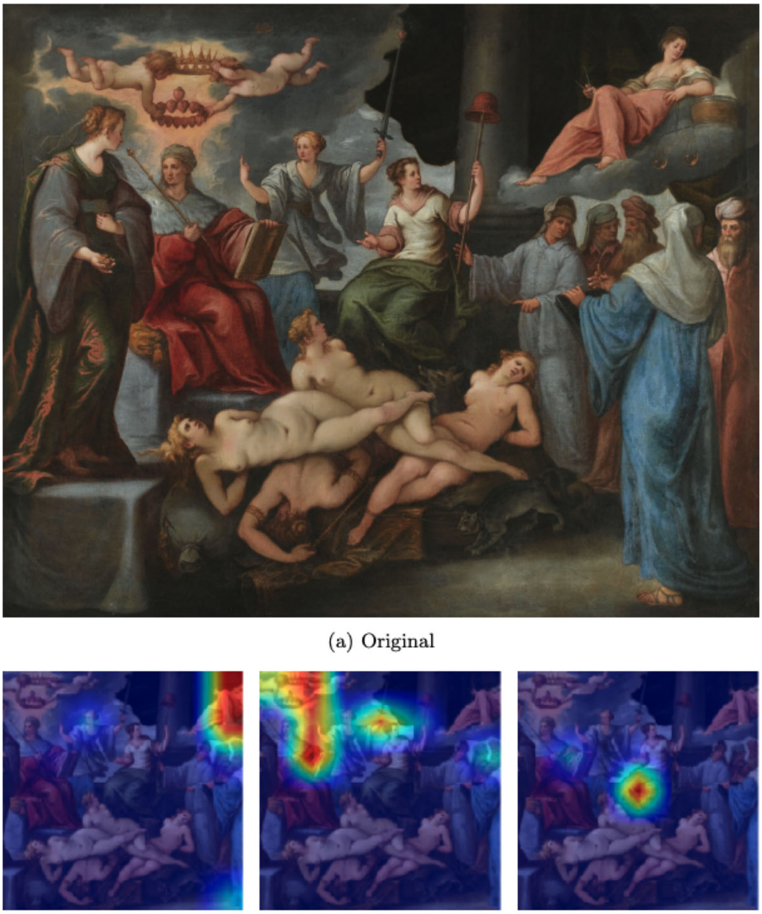
5.2 Results

Class activation mapping

We rely on GradCAM++ to experiment in identifying which parts of certain images are valuable from the point of view of the fine-tuned classification models. The saliency maps for images of interest previously unseen by the model are done by activating specific classes in our best model, VGG-16 finetuned for 1000 epochs. We present some results on images previously unseen by the model in Figs. 7 and 8). The latter figure presents the resulting heatmaps for *freedom* on Delacroix's iconic painting as well as on two derivative works inspired by it, which re-interpret Delacroix's within the context of Hong Kong protests.[17]

[16] https://github.com/CompVis/stable-diffusion

[17] "Liberty Leading the People of Hong Kong" by Frederic Bussiere, group exhibition "The Art of Resistance", Kong Art Space, Hong Kong, 2019, Digital collage, printed on canvas. 90 x 90 cm https://www.behance.net/gallery/90838377/Liberty-Leading-the-People-of-Hong-Kong-collage and "Our Vantage" by Harcourt Romanticist https://www.instagram.com/p/B2EQ1FPnDh1/



(a) Original

(b) Comfort

(c) Freedom

(d) Safety

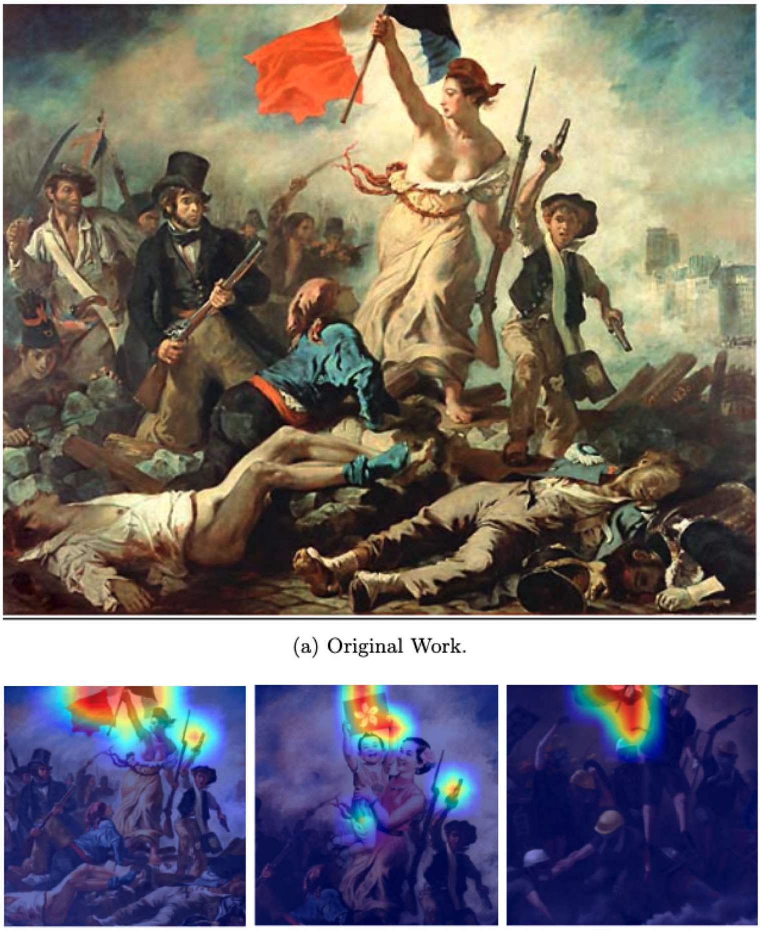
Fig. 7 GradCAM++ for three different classes computed using VGG-16 finetuned for 1000 epochs on *Triumph of the Virtues over the Vices* painting by *Paolo Fiammingo*, circa 1592. Oil on canvas, dimensions 16.5 × 221 cm (6.4 × 87 in). Image sourced from Wikimedia Commons, originally from Sotheby's auction in London on 6 July 2011

Activation maximization

In this section, we present the results of applying the same logic of Offert (2019) to create (activation maximization) feature visualizations for each class of the models trained on ARTstract. We present the class feature visualizations for optimizing the activation for each of the 8 target classes for two models: the VGG model fine-tuned for 100 epochs (Fig. 9a), and the VGG model fine-tuned for 1000 epochs (Fig. 9b).

Stable diffusion-activation maximization

With the stable diffusion-based image generator model, we synthesized realistic images from the ones obtained using the AM method. Example results from both



(a) Original Work.

(b) Heatmap for *freedom* in Original Work.

(c) Heatmap for *freedom* in Derivative Work 1.

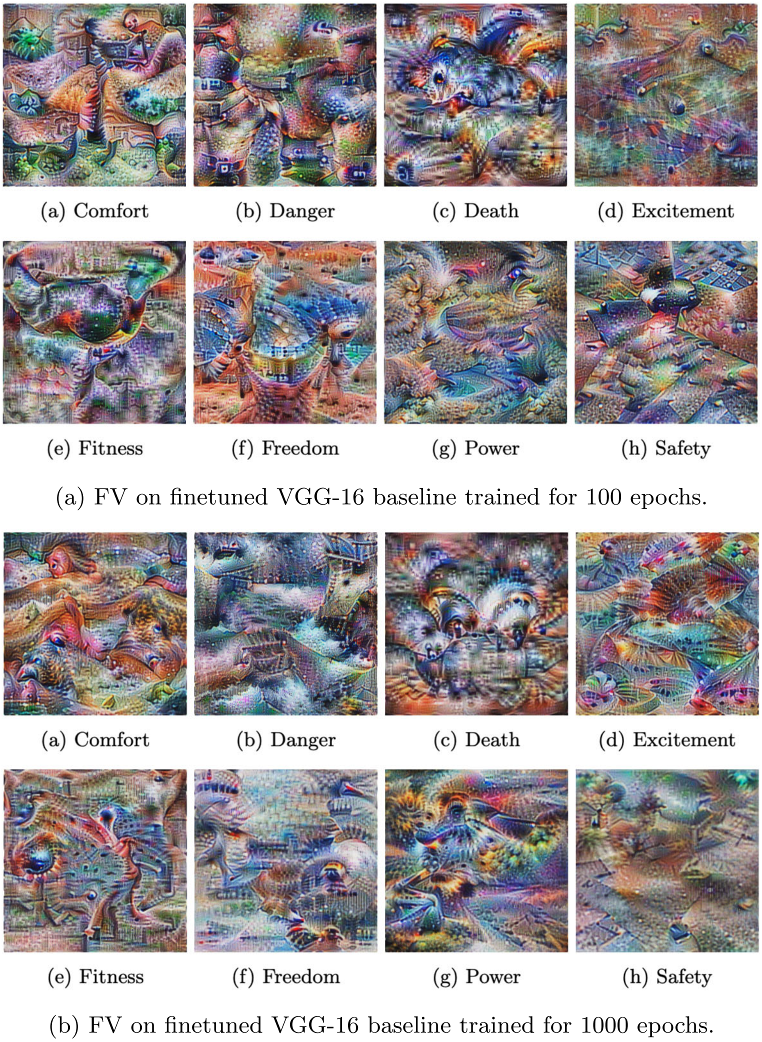
(d) Heatmap for *freedom* in Derivative Work 2.

Fig. 8 Top: *Liberty Leading the People*, oil painting by Eugène Delacroix, 1830, dimensions 260 cm × 325 cm, Louvre, Paris; image sourced from Wikimedia Commons. Bottom: GradCAM++ computed for *freedom* on (1) the original painting, and two (2-3) derivative paintings. The three heatmaps show similar activation areas for the class of *freedom*

experimental settings are hereby presented: for *death* and *freedom* in Fig. 10, and for *fitness* and *comfort* in Fig. 11.)

6 Discussion

Abstract concepts stand as high-level semantic units within the realm of computer vision, spotlighting the limitations of binary thinking. Laden with ambiguity, subjectivity, and context-dependency, abstract concepts defy facile classification. The



(a) Comfort (b) Danger (c) Death (d) Excitement

(e) Fitness (f) Freedom (g) Power (h) Safety

(a) FV on finetuned VGG-16 baseline trained for 100 epochs.

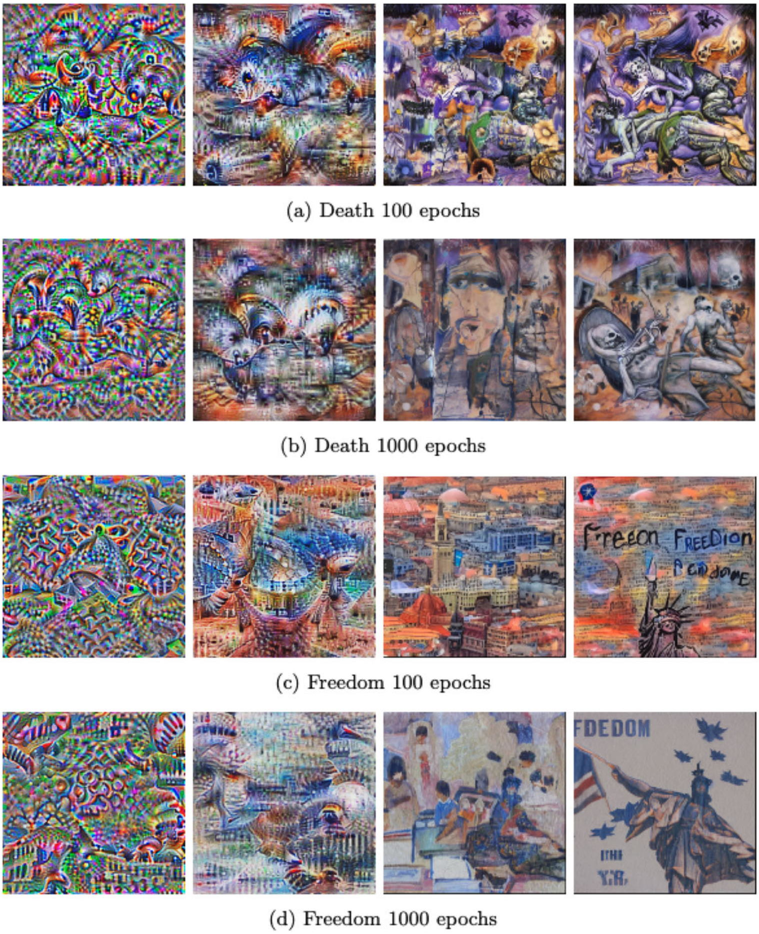
(a) Comfort (b) Danger (c) Death (d) Excitement

(e) Fitness (f) Freedom (g) Power (h) Safety

(b) FV on finetuned VGG-16 baseline trained for 1000 epochs.

Fig. 9 Feature visualizations for each of the eight target classes using VGG-16 finetuned for (a) 100 and (b) 1000 epochs

diverse meanings these concepts embody within varying cultural and historical contexts further compound the intricacies of labelling. In stark contrast to problems marked by well-defined categories, such as tame problems, high-level computer vision tasks encapsulate the very essence of wicked problems, demanding collective thought and an integrative approach to navigate their intricate landscape. This conceptualization of wicked versus tame problems draws inspiration from the realm of social planning and political science, where wicked problems, characterized by multidimensionality, multiculturalism, and strong cultural dimensions, necessitate nuanced and collaborative solutions (Rittel, 1967).



(a) Death 100 epochs

(b) Death 1000 epochs

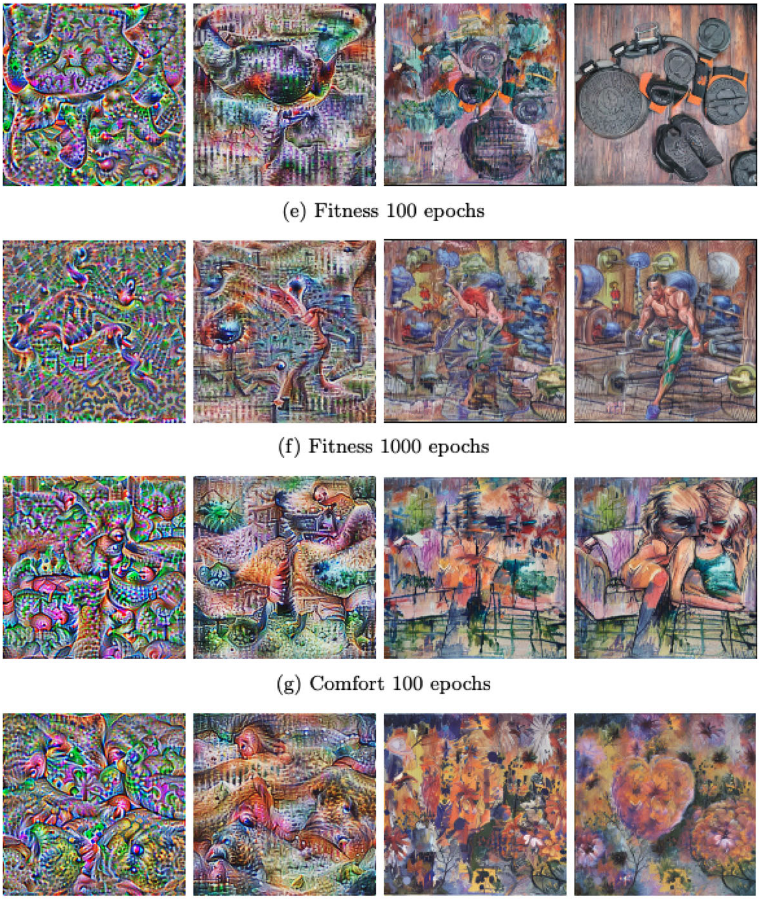
(c) Freedom 100 epochs

(d) Freedom 1000 epochs

Fig. 10 Denoising of feature visualizations (FV) of *death* (first row: 100 epochs model, second row: 1000 epochs model) and *freedom* (third row: 100 epochs model, fourth row: 1000 epochs model). The FVs have been denoised with different levels of regularization and using Stable Diffusion (SD) and prompt-guided SD methods. The leftmost FV is minimally regularized, followed by a regularized FV, and then SD denoised FV. The rightmost FV shows a regularized FV denoised with prompt-guided SD

6.1 ARTstract coverage

The ARTstract dataset is a new resource for cultural heritage and computer vision research, particularly in digital humanities projects. However, some key limitations of the dataset are related to the inherent nature of attempting to create a clear definition of an abstract concept. Furthermore, as we chose to reuse and combine existing datasets, ARTstract is bound by choices made by the creators of these datasets. We acknowledge that abstract concepts are inherently culturally motivated, this is perhaps most visible in the ADVISE dataset as the advertising domain is highly culturally contextualised. However, also in Artpedia and the Tate Gallery Western (European)



(e) Fitness 100 epochs

(f) Fitness 1000 epochs

(g) Comfort 100 epochs

(h) Comfort 1000 epochs

Fig. 11 Denoising of feature visualizations of *fitness* (first row: 100 epochs model, second row: 1000 epochs model) and *comfort* (third row: 100 epochs model, fourth row: 1000 epochs model). The FVs have been denoised with different levels of regularization and using Stable Diffusion (SD) and prompt-guided SD methods. The leftmost FV is minimally regularized, followed by a regularized FV, and then SD denoised FV. The rightmost FV shows a regularized FV denoised with prompt-guided SD

art makes up a larger proportion which has a direct impact on ARTstract's coverage and representation. Within these artworks, certain themes and symbols form a shared conceptual grounding to their intended audiences (Panofsky & Drechsel, 1955), therefore these datasets are suited to use for our purpose, with the caveat that they represent a particular context. The coverage limitation within ARTstract stems partly from the intrinsic challenge of cultural bias when assigning labels and meanings to images, particularly concerning abstract concepts. Moving forward, it is imperative to expand coverage and enhance the explicit contextualization of labels, dispelling the notion of their objectivity.

Despite its limitations, ARTstract fills a significant gap by providing much-needed data for tasks related to AC image classification. Critically, we believe that the richness and diversity of the ARTstract dataset provide a unique opportunity for exploring and experimenting with explainable CV methods. The dataset offers a testing ground for existing and novel explainable CV methods, demonstrating the potential of combining technical methods with hermeneutic work to develop interpretable systems. Moreover, the significance of the ARTstract dataset goes beyond its value as a resource for cultural heritage and CV research. The evocation of abstract concepts is complex, subjective, and culturally variant, and as such, we hope that the development of this dataset can be a source of inspiration to expand it with more complex, situated, and multicultural perspectives.

6.2 Abstract concept image classification performance

The results of this study bring to the fore the wicked nature of the problem of automatically detecting abstract concepts within computer vision. The proposed AC image classification baselines show relatively low performances when compared to other CV tasks on art images, such as style, genre, or artist classification (Tan et al., 2016; Cetinic et al., 2018). The results of Table 2, however, are similar to the results obtained by models that use a similar amount of images on a radically different set of labels compared to ImageNet (Ng et al., 2015). The complexity of detecting abstract concepts might hence stem from the relatively open definition of each abstract concept, which does not explicitly account for their polysemy and association to vastly varied visual data (as seen in the example of *danger* in Fig. 4). The shallow representation obtained using a CNN-based method is not able to generalise enough to capture the ambiguities of such definitions. The confusion matrices in Fig. 6, indeed, reveal potential co-occurrences of abstract concepts that prompt further investigations, such as in the case of the *power* class.

The decision to train models for abstract concept image classification, even when expecting low performances, presents a critical consideration in our study. The results obtained in Table 2 of Section 4 highlight how current existing models are not well suited to classify AC yet. Even though more complex training procedures can be employed (Kandel & Castelli, 2020) (e.g. training only a subset of the total number of convolutional layers) or more powerful models can be used, such as ViT (Zhai et al., 2022), we argue that the intra-class variance displayed by ARTstract hardly allows any significant improvement from a quantitative point of view (Benz et al., 2020; Shirali & Hardt, 2023). The use of pre-trained models, such as those from ImageNet, offers a promising foundation by leveraging their learned visual features. Training a model for AC classification allows us to investigate to which extent low-level perceptual features can be used on this task. This allows us to better understand which insights, if any, from these methods can be used to more effectively deal with the task. The low performances can be seen as a reminder of the intrinsic complexity of this wicked challenge, underscoring the need for interdisciplinary, collaborative thinking and an adaptive, iterative approach to foster a deeper understanding of complex visual concepts.

6.3 Insights from the explainability experiments

This recognition of the wicked nature of automatically detecting abstract concepts underscores the pressing need for interpretability, collaborative inquiry, and nuanced approaches to tackle the challenges posed by these elusive and intricate visual phenomena. When confronted with wicked problems, traditional black-box solutions fall short in providing meaningful insights. The inherent complexity, contextual dependencies, and subjective nature of abstract concepts demand a level of transparency and explainability in computer vision methods. Interpretability becomes paramount not only for improving model performance but also for gaining insights into the underlying reasons for model decisions. By uncovering the visual cues and features that contribute to classification outcomes, interpretability facilitates informed refinement and adaptation of algorithms. This intertwining of wicked problems with the demand for explainability forms the cornerstone of our study's motivation, driving us to explore the boundaries of computer vision in addressing intricate and inherently human challenges.

6.3.1 Insights from traditional explainability: CAM

We experimented with a CAM-based method as it is the most well-known explainability technique to uncover the decision-making mechanisms of CNN-based models. These insights were pivotal in shedding light on the complex processes underlying our models' decisions. To explore the reasons behind our models' classification of unseen images, we employed GradCAM++ to pinpoint valuable image regions. Using the VGG-16 model finetuned over 1000 epochs, we activated specific classes for previously unencountered images. This approach provided valuable perspectives into the model's decision boundaries. For example, using the concept of *freedom* presents a distinct situation where intriguing connections are revealed. Figure 8 displays the results for the class of *freedom* in Eugène Delacroix's painting "Liberty Leading the People" and two derivative works inspired by it. In this example, both the original work and the two derivative works localise the *freedom* class in the same area (i.e. where the flags and raised hands are located). We can hence discern how the model identifies specific visual cues that evoke the abstract concept. For instance, the emphasis on flags within the tested images suggests a connection between the concept of *freedom* and symbols of nationhood or political expression. This nuanced exploration of the model's decision-making process underscores the role of explainability techniques in unravelling the intricate relationships between abstract concepts and their visual manifestations.

Another aspect that emphasises the usefulness of such a technique is its consistent application on the same image, but to localize important regions for different classes. Figure 7 presents the results on three different classes in the painting *"Triumph of the Virtues over the Vices"*. These experiments identify parts of the image in which the model focuses for the image's classification as a selected class. For instance, *comfort* is localized near the figure sitting in a relaxed position on a comfortable couch. In contrast, *freedom* is concentrated around the area of the painting with angels, clouds, and a raised sword. These findings suggest biases in the ARTstract dataset, potentially stemming from *freedom*-tagged images being biased towards images depicting elements like raised swords or flying agents such as birds or angels.

Overall, these results showcase the effectiveness of CAM-based methods in identifying valuable regions in images for classification models, thereby highlighting potential biases in the dataset and providing insights into how the model perceives and processes images. However, the results also underscore that while CNNs are aware of statistical correlations, these correlations may not always align with human perspectives. Despite providing valuable insights into classification processes and the identification of abstract concepts in images, the shallow representation achieved by the model can yield false evidence. Furthermore, the lack of robustness, particularly against adversarial attacks, poses a significant concern for the interpretability of classifiers (Akhtar et al., 2021). In conclusion, these findings stress the need for further research to enhance the accuracy and robustness of classification models when addressing abstract concepts in the realm of art.

6.3.2 AM: distributed reality, perceptual bias, and feature visualization

We see the regularized feature visualizations for each of our eight abstract concept targets (shown in Fig. 9) as examples of how distributed reality (in terms of manifestations and perspectives) can get collapsed into one 2D image. When they are learned and represented by a CNN, concepts are “dissolved”, or “entangled”, losing their spatial coherence, and thus “it is no surprise that feature visualization images will reflect different manifestations of, and perspectives on, an object, akin to Cubist paintings” (Offert & Bell, 2021, pg. 1301). We see abstract concepts as a prime example of how visual concepts get dissolved in ways that are practically unintelligible to humans. Additionally, the fact that the images in the referenced figure are regularized means that we already introduced a syntactic bias so as to guide the manifestations into a textural landscape closer to what we visually comprehend. With this syntactic optimization, most of the FVs in Fig. 9 are still relatively humanly incomprehensible (no noticeable objects or otherwise *legible* items are very visible). An exception may be

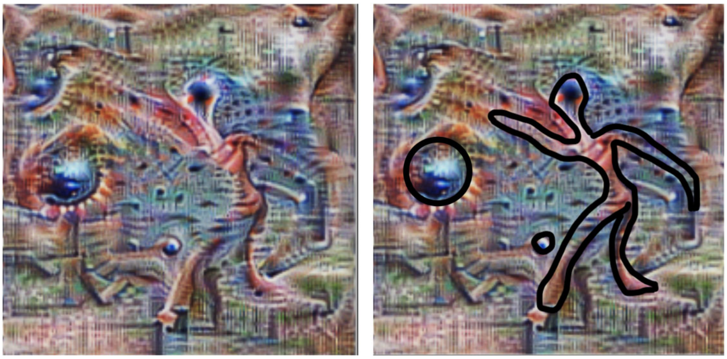


Fig. 12 The feature visualization (FV) of the *fitness* class using VGG-16 finetuned for 1000 epochs (left) resembles a human figure playing with a sort of ball (right)

the case of the FV of *fitness* for the model finetuned for 1000 epochs, in which certain edges seem to resemble a human figure playing with some sort of ball (see Fig. 12).

6.3.3 SD-AM: Denoising and the Big Trade-Off

Because to move from the entangled state to a state of higher semantic interpretability for humans requires introducing more constraints, in this work we decided to combine the feature visualizations produced with the AM method with Stable Diffusion (SD-AM) (see Fig. 13). Despite some work on the generation of counter-factual explanations (Jeanneret et al., 2022; Zemni et al., 2023), the investigation of diffusion models trained on large-scale data, such as Stable Diffusion, as tools that allow a better understanding of classification models is a novel approach. While past attempts, such as combining AM with GAN models (Nguyen et al., 2016) have shown promising results, we argue that the intrinsic dependency of GANs on a classification model (the discriminator) can potentially bias the generation process toward results that are harder to automatically classify for syntactic reasons (i.e. the color distribution) rather than for semantic reasons. This is especially true for abstract concepts, where the difference between classes, for instance *danger* and *power*, depends on the semantic content of the image alongside the interpretation of the abstract concept that the network inductively learns. The use of diffusion models overcomes this limitation by design. An image is not generated in order to fool a discriminator, but rather with the aim of removing the noise that makes an input image hard to interpret for humans. While the experiments on FV (Fig. 9) provide little insight into the perception of an AC by the model, the denoised version of such images, especially those that are textual prompt

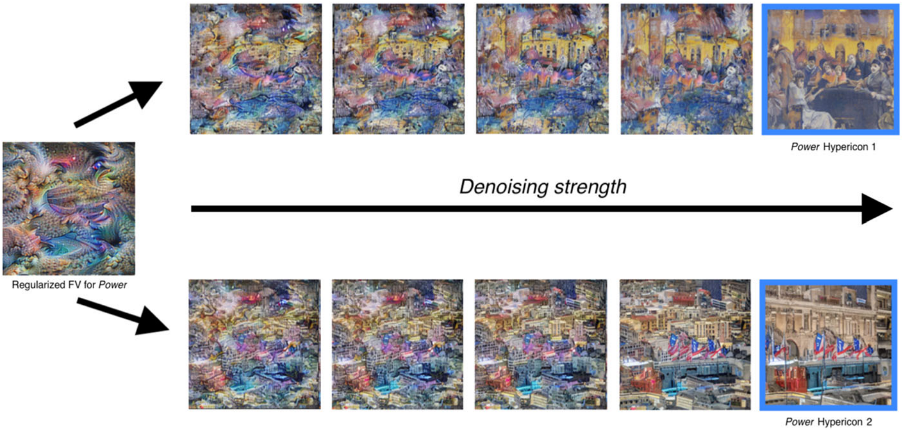


Fig. 13 Starting with the regularized FV for the *power* class (derived from VGG-16 trained for 100 epochs), the progression from left to right illustrates a gradual escalation in denoising strength, resulting in images with enhanced human interpretability. This procedure was performed twice, yielding distinct SD-AM "hypericons" for the *power* category (distinguished by blue borders), both originating from the same initial regularized FV. It is noteworthy that the denoising process was executed **independently of any textual prompt**, thereby ensuring that the process remained entirely oblivious to the correlation between the FV earmarked for denoising and its status as representing "power"

guided, (Figs. 10 and 11) let prototypical versions of an AC emerge. For instance, both the models, finetuned for 100 and 1000 epochs, represent *freedom* by means of an icon resembling the Statue of Liberty. Similarly, for the *fitness* class, hypericon images resembling a gym are obtained from the noisy FV.

However, it is important to note that the images produced are confined to the latent space of the specific diffusion model employed, similar to the one argued regarding GANs (Offert & Bell, 2021). These images do not reflect the perceptual topology of the analysed CNN, but they rather replace the elements that are hard to interpret with what the model perceives as human-like, essentially filling the gap between AM images and interpretable hypericons. As such, SD-AM is to be taken as an explorative approach towards easier interpretation of AM images. Further research is required to investigate whether this approach can be directly incorporated in the extraction of saliency maps, to obtain human interpretable results while minimising the influence of SD.

Perhaps the most critical contribution of the studies that inspired our work (Offert, 2019; Offert & Bell, 2021) is their discussion of the problem of perceptual bias in machine vision systems, which can only be overcome by shifting toward different biases. As they discuss, any constraint added to the optimization process for feature visualizations moves the images further away from showing the actual perceptual topology of a CNN, unveiling the trade-off between representational capacity and legibility of feature visualization images. Their work highlights that feature visualization is one way to achieve forced legibility but also presents a dilemma that the representational capacity of feature visualization images is inversely proportional to their legibility. Feature visualizations that show “something” are further removed from the actual perceptual topology of the machine vision system than feature visualizations that show “nothing.”

While keeping in mind this trade-off, they also note that we can treat resulting hypericons as valuable tools for visual interpretability. Hypericons, such as the ones presented in Fig. 10 and discussed in the previous section can be used in combination with the original datasets (in this case, ARTstract) to enable the identification of interesting patterns, especially when treating them as (Mitchell, 1995, p. 49) suggests:

> “The metapicture is a piece of moveable cultural apparatus, one which may serve a marginal role as illustrative device or central role as a kind of summary image, what I have called a ‘hypericon’ that encapsulates an entire episteme, a theory of knowledge.”

As such, in addition to aiding the understanding of the perceptual topologies of CV models, feature visualization images can be studied as concrete representations of cultural knowledge defined by the lenses and tags fed into the CV systems. We hope that our analyses in this case study can function as evidence that exactly this “subjective” nature of feature visualization images is what can make “visual explainability useful in computer vision for art” Offert (2019).

A crucial point is that the SD-AM can lead to relevant and semantically meaningful hypericons even without any textual prompt, i.e., without being guided or biased by the class corresponding to the input AM image. For example, in Fig. 13, we present results of applying promptless SD-AM to obtain hypericons related to the *power* class. We denoised the extracted and regularized AM by gradually increasing the intensity (weight). By increasing the intensity of the denoising process, we are able to control the number of transformations applied by SD to obtain an image that is perceived (by the model) as closer to its original training data. Critically, as seen in Fig. 14, this process effectively converges towards more human-intelligible hypericons, which resemble real instances of artworks from the corresponding class present in the original ARTstract dataset both visually and semantically.

6.4 Hyperpop hypericons

The proliferation of images in modern mass media has reached unprecedented levels, with social media platforms alone hosting billions of images every day, resulting in a visually abundant contemporary culture where individuals are bombarded with heterogeneous visual data. This phenomenon characterizes the post-modern era, where users are overwhelmed by an abundance of information that is not curated (Jansson & Hracs, 2018), making it increasingly important to become "lookers" in addition to being readers (Smits, 2022). The rise of hyperpop and meme culture, favored by the newest generation of technology users searching for sense amidst the chaos, is symptomatic of the current historical moment. As Vassar (2020) suggests, hyperpop serves as an attempt, "suited to the psyche of the six-hours-of-screen-time-a-day individual", to strive to find some semblance of meaning amidst the disarray, by compiling a vast field of disparate meanings until they reach some semblance of accord.

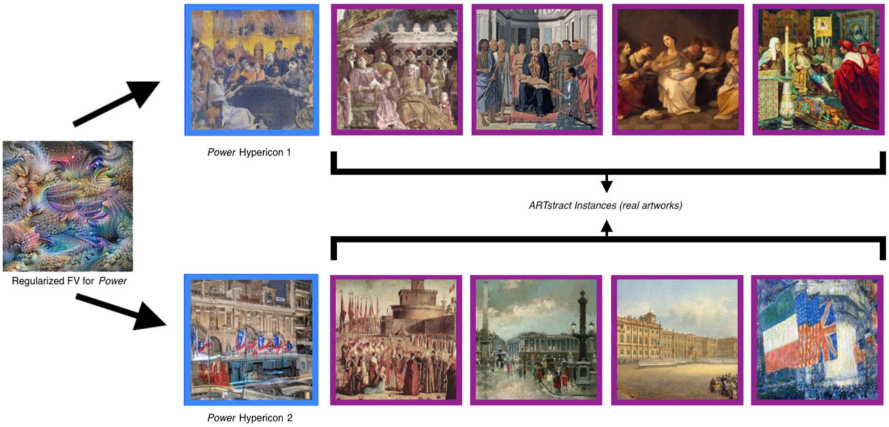
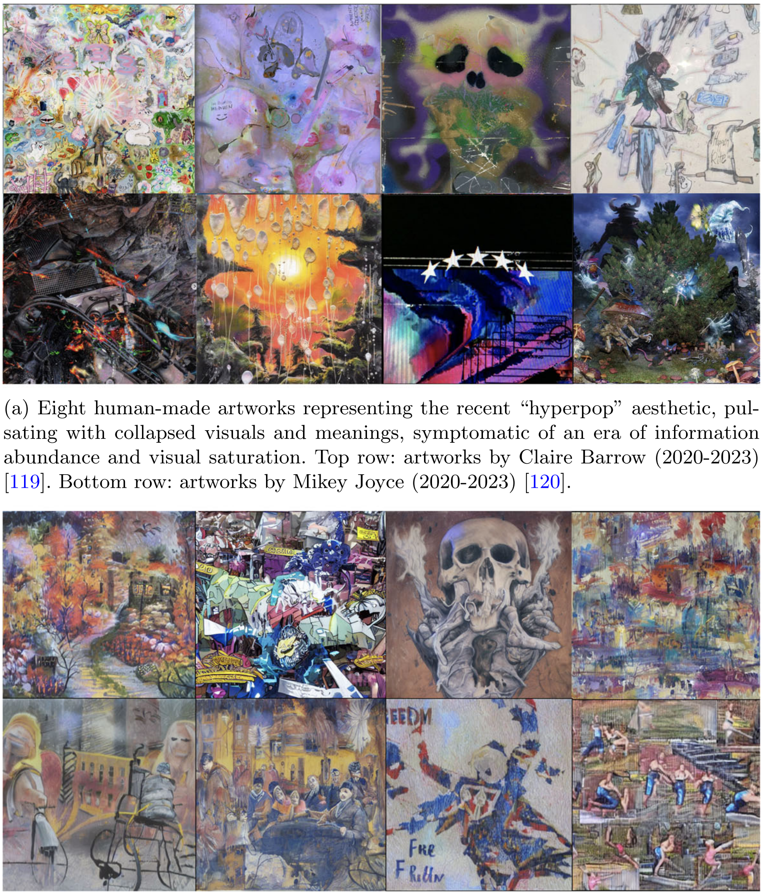


Fig. 14 Comparison of the synthetic SD-AM "hypericons" for the *power* class (with blue borders) with manually selected with real instances from ARTstract (with purple borders). These real images from ARTstract are tagged with *power*, they were selected because of their visual and semantic similarity to the hypericons

Intriguingly, an uncanny resemblance between the SD-AM hypericons and hyperpop artworks surfaces, as illustrated in Fig. 15. Within this visual dialogue, a profound parallel emerges as both categories exude dissolved yet intricately collapsed visuals. These striking visuals offer a blend of object fragments and hues, all while boundaries remain indistinct. This parallel beckons us to explore the intersections of aesthetics and cognition. Intriguing questions arise—does the act of meaning collapse within the hypericons mirror the cognitive underpinnings of the hyperpop aesthetic? Could it signify a convergence of overstimulated sensibilities, reminiscent of the torrent of data processed by our models? In an analogous manner, just as our hypericons weave significance from intricate data, the hyperpop aesthetic might mirror the cognitive fab-



(a) Eight human-made artworks representing the recent "hyperpop" aesthetic, pulsating with collapsed visuals and meanings, symptomatic of an era of information abundance and visual saturation. Top row: artworks by Claire Barrow (2020-2023) [119]. Bottom row: artworks by Mikey Joyce (2020-2023) [120].

(b) Eight SD-AM hypericons crafted for each of the 8 abstact concept classes, emblematic of our work, revealing a surprising resemblance to the aesthetic of hyperpop artworks.

Fig. 15 Visual convergence of (15a) hyperpop aesthetics and (15b) SD-AM hypericons, a juxtaposition that invites contemplation on the parallels between between the rapid pace of modern media consumption and the massive data flow into deep learning models

ric of contemporary generations, offering a fresh lens for interpreting the avalanche of visual content.

6.5 Explainability lessons for DH

We can assume that detecting and correcting bias of computer vision systems in the context of Cultural Heritage (CH) will mostly happen in a *post-hoc* manner, i.e., after a system has been deployed in real-world situations. This is due to the fact that many models based on similar patterns have already been used in real-world applications, especially in the digital humanities. We believe that the integration of interpretability into CV-based systems in the cultural heritage (CH) field has not received enough attention. This study stands as proof that digital humanities (DH) initiatives can act as valuable arenas for both probing the limits of established CV explanation techniques and pioneering novel methodologies. DH projects, with their interdisciplinary focus and emphasis on interpretation, offer a unique opportunity to combine technical methods with hermeneutic work to develop systems that are interpretable-by-design. We envision our work as one of many DH projects that can contribute to the broader development of more transparent and understandable computer vision systems. With their diagnostic capability, the tools developed in XAI are exciting both to the technical disciplines for improving the systems they develop, and to fields such as Digital Humanities with alternative paths for thinking about the kind of work they do (e.g. by interrogating through explainable methods the way that a system has classified certain cultural objects) (Berry, 2021).

7 Future directions

There are multiple further directions that can be taken with this work. First, recognizing and rectifying the cultural bias embedded in assigning meaning to visual data through labels is a crucial step that should be formally acknowledged and tackled in a machine-readable manner (Pandiani & Presutti, 2022). Future research should focus on addressing and explicitly tracking diverse cultural contexts and their associated abstract concepts, offering alternative perspectives and insights on these concepts. Another of the main current limitations of the ARTstract dataset is its size compared to other art datasets, such as WikiArt (Mohammad & Kiritchenko, 2018), Web Gallery of Art (Cetinic et al., 2018) or the TICC Printmaking Dataset (van Noord & Postma, 2017). This prevents the effective use of models that requires large sets of input images, such as current state-of-the-art methods in the image classification tasks (Chen et al., 2021). A possible way to overcome this issue is to extend ARTstract to include more and other types of artistic creations. An interesting approach is that of extending it by means of automatic tagging for example by using techniques from the natural language processing domain (Lin, 2022), such as the extraction of AC from descriptive sentences using topic modelling (Kherwa & Bansal, 2020) or linguistic frames (Presutti et al., 2012). Similarly, we plan on improving the definition of the abstract concept classes, by expanding clusters and triggers and providing richer descriptions.

This involves the alignment with resources like WordNet (Miller, 1998) and BabelNet (Navigli & Ponzetto, 2012) to provide a more comprehensive definition of clusters as well as relevant ontologies (Baldoni et al., 2012; Bertola & Patti, 2016) to provide semantically richer descriptions that allow the integration with other relevant artwork information.

A different approach to improve AC image classification performance is to further pre-train models, initially trained on ImageNet, on the other available art datasets. Specialisation in AC detection can then be achieved by fine-tuning the final convolutional layers on ARTstract (Kandel & Castelli, 2020). Additionally, other vision architectures can be explored, such as ViT (Zhai et al., 2022), as well as hybrid methods using Knowledge Graphs (KG) as background knowledge (Marino et al., 2017; Zhang et al., 2019). Further directions include running attention in an unsupervised way to test whether it can locate the areas of images that are most important for evoking an AC, which can be used to later designate bounding boxes (Ibrahim & Shafiq, 2023). The experiments performed with GradCAM++, such as those of Fig. 8, suggest that it might be possible to detect specific regions on which an AC can be more prominently identified. We will further investigate this aspect by systematically evaluating the regions identified using other CAM-based methods (Nguyen et al., 2019) as well as methods that achieve state-of-the-art results in the semantic segmentation task (Mo et al., 2022).

Improving AC classification explainability could include a plethora of further directions: conducting color palette analysis, identifying co-occurring objects, actions, and colors through object detection, and exploring pose detection. Other visual cues such as texture, shape, and style can be investigated. Additionally, we could expand experiments on adversarial training, style detection, and the synthetic image using guided diffusion can be explored, which may involve exploring the use of other generative models such as GANs. Another potential avenue for further research is to investigate the use of KG and empirical semantics to enhance explainability. This may involve developing new methods for incorporating KG and semantic information into CV systems, as well as exploring the potential of explainability techniques such as attention, saliency maps, and semantic segmentation in the context of KG-based CV. Overall, we can further our work by exploring other datasets, techniques, and methodologies from related fields such as art history and cultural studies.

8 Conclusion

In our pursuit to address the intricate challenge of automatically classifying images based on evoked abstract concepts (ACs), we have introduced the novel ARTstract dataset as a lens through which we delve into the realm of explainability. This work unravels the role of Convolutional Neural Networks (CNNs) in such complex high-level visual tasks, but also establishes benchmark model performances for AC image classification within the ARTstract context. Additionally, we combine traditional and novel explainability techniques to better understand model behavior and predictions. With SD-AM, by harmonizing activation maps (AM) with diffusion models, we create synthetic “hypericons” that compellingly visualize the profound transformation of

AC meanings as captured by deep networks into singular images. Our study resonates with the burgeoning demand for interpretability in computer vision systems, especially within the cultural heritage domain and the realm of socio-cultural-cognitive visual understanding. We accentuate the significance of recognizing biases and forging connections between the technical and humanistic dimensions, advocating for unconventional pathways to extend hermeneutics. In conclusion, our article calls for the explicit integration of explainability into the fabric of CV-based systems that attempt to address high-level visual challenges. This integration is vital to ensure the dependability and credibility of these systems in the evolving landscape of art, culture, and technology. It beckons us to challenge the binary boundaries that prevail in CV, advocating for a more holistic, humanistic, and ethical perspective.

Acknowledgements We wish to express our sincere gratitude to Nanne van Noord and Melvin Wevers for their invaluable contributions to our research endeavor. Their insightful suggestions, innovative ideas, and thoughtful recommendations significantly influenced both the conceptual and technical aspects of our work. Their generous support and guidance were crucial in shaping the direction of our study. We would also like to extend our appreciation to Stefano Zingaro, who provided us with technical assistance in fine-tuning the computer vision models used in our analyses. Finally, we express our gratitude to Valentina Carriero for her idea in naming the ARTstract dataset.

Author Contributions D.S. Martinez Pandiani: conceptualization of the research and explainability experiments, dataset creation and training of baseline models, developed the methodology, conducted investigations, created software, validated the results, drafted and reviewed the manuscript. N Lazzari: improved upon the explainability experiments, and played a significant role by contributing to the software, validating the results, and drafting and reviewing the manuscript. M. van Erp contributed to the original draft, reviewing the manuscript, and provided supervision. V. Presutti provided further supervision.

Funding We acknowledge the support and resources provided by the Marco Polo Initiative (2022) from the University of Bologna, which allowed D.S. Martinez Pandiani to conduct a period of research at the KNAW Humanities Cluster in Amsterdam.

Availability of data and materials All data and algorithms introduced in this study are openly accessible and available for further research at the following repository: https://github.com/delfimpandiani/Hypericons-x-AbstractConcepts.

Declarations

Ethical Approval Not applicable.

Competing interests We declare that there are no competing interests, including any financial or personal relationships that could potentially influence the results or interpretation of this work.

References

Abgaz, Y., Rocha Souza, R., Methuku, J., Koch, G., & Dorn, A. (2021). A methodology for semantic enrichment of cultural heritage images using Artificial Intelligence technologies. *J Imaging,* 7(8), 121. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/jimaging7080121.

Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., & Guibas, L.J. (2021). ArtEmis: Affective Language for Visual Art. In: IEEE Conference on computer vision and pattern recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE pp. 11569–11579. Available from: https://openaccess.thecvf.com/content/CVPR2021/html/Achlioptas_ArtEmis_Affective_Language_for_Visual_Art_CVPR_2021_paper.html.

Ahres, Y., & Volk, N. (2016) Abstract Concept & Emotion Detection in Tagged Images with CNNs. Unpublished Report, accessed from http://cs231nstanfordedu/reports/2016/pdfs/008_Reportpdf. p. 8.

Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. *IEEE Access, 9*, 155161–155196. https://doi.org/10.1109/ACCESS.2021.3127960

Baldoni, M., Baroglio, C., Patti, V., & Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale, 6*(1), 41–54. https://doi.org/10.3233/IA-2012-0028

Barthes, R. (1980). Camera Lucida: Reflections on Photography, trans. R. Howard, New York: Hill & Wang. Orig. La Chambre Claire, Note sur la Photographie.

Bekkouch, I.E.I., Eyharabide, V., & Billiet, F. (2021). Dual Training for Transfer Learning: Application on Medieval Studies. In: 2021 International joint conference on neural networks (IJCNN). IEEE pp. 1–8.

Benz, P., Zhang, C., Karjauv, A., & Kweon, I.S. (2020). Robustness May Be at Odds with Fairness: An Empirical Study on Class-wise Accuracy. In L. Bertinetto, J.F. Henriques, S. Albanie, M. Paganini, & G. Varol (Eds.), NeurIPS 2020 Workshop on pre-registration in machine learning, 11 December 2020, Virtual Event. vol. 148 of Proceedings of Machine Learning Research. PMLR pp. 325–342. Available from: http://proceedings.mlr.press/v148/benz21a.html.

Berry, D. (2021). at MIT Libraries DH, B. David (Ed.), The explainability turn and Digital Humanities. MIT Libraries Youtube. Available from: https://www.youtube.com/watch?v=cvHwiBD_EHs.

Berry, D.M. (2022). AI, Ethics, and Digital Humanities. The Bloomsbury Handbook to the Digital Humanities. p. 445.

Bertola, F., & Patti, V. (2016). Ontology-based affective models to organize artworks in the social semantic web. *Inf Process Manag, 52*(1), 139–162. https://doi.org/10.1016/j.ipm.2015.10.003

Bevan, A. (2015). The data deluge. *Antiquity, 89*(348), 1473–1484.

Borghi, A.M., & Binkofski, F. (2014). Words as social tools: An embodied view on abstract concepts. vol. 2. Springer

Boyd Davis, S., Vane, O., & Kräutli, F. (2021). Can I believe what I see? Data visualization and trust in the humanities. *Interdisciplinary Science Reviews, 46*(4), 522–546.

Brigato, L., Barz, B., Iocchi, L., & Denzler, J. (2022). Image Classification With Small Datasets: Overview and Benchmark. *IEEE Access, 10*, 49233–49250. https://doi.org/10.1109/ACCESS.2022.3172939

Bykov, K., Hedström, A., Nakajima, S., & Höhne, M.M.C. (2022). NoiseGrad - Enhancing Explanations by Introducing Stochasticity to Model Weights. In: Thirty-Sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, The twelveth symposium on educational advances in artificial intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press pp. 6132–6140. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/20561.

Calabrese, A., Bevilacqua, M., & Navigli, R. (2020). Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics pp. 4680–4686. Available from: https://aclanthology.org/2020.acl-main.425.

Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*. 18(2), 1–22

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning Convolutional Neural Networks for fine art classification. *Expert Syst Appl, 114*, 107–118. https://doi.org/10.1016/j.eswa.2018.07.026

Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V.N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Winter conference on applications of computer vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018. IEEE Computer Society. pp. 839–847. Available from: https://doi.org/10.1109/WACV.2018.00097.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of Image Classification Algorithms Based on Convolutional Neural Networks. *Remote Sens, 13*(22), 4712. https://doi.org/10.3390/rs13224712

Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of the ACM international conference on image and video retrieval. CIVR '09. New York, NY, USA: Association for Computing Machinery. pp. 1–9. Available from: https://doi.org/10.1145/1646396.1646452.

Cohen, J. N. M., & Mihailidis, P. (2013). *Exploring Curation as a core competency in digital and media literacy education* (p. 4). Faculty Works: Digital Humanities & New Media.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee pp. 248–255.

Deutch, D., Malik, T., & Chapman, A. (2022). Theory and Practice of Provenance. In: Proceedings of the 2022 International Conference on Management of Data pp. 2544–2545.

Dhariwal, P., & Nichol, A.Q. (2021). Diffusion Models Beat GANs on Image Synthesis. In M. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang, & J.W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021,* NeurIPS 2021, December 6-14, virtual; 2021. p. 8780–8794. Available from: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal,* 1341.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal, 1341*(3), 1.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., & Schöch, C., et al. (2017). Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities. 32(suppl_2):ii4–ii16.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In: 31st British machine vision conference 2020, BMVC, virtual event, UK, September 7-10, 2020. BMVA Press; 2020. Available from: https://www.bmvc2020-conference.com/assets/papers/0631.pdf.

Gella, S., Elliott, D., & Keller, F. (2019) Cross-lingual Visual Verb Sense Disambiguation. arXiv:1904.05092 [cs]. arXiv:1904.05092

Gella, S., Lapata, M., & Keller, F. (2016) Unsupervised Visual Sense Disambiguation for Verbs using Multimodal Embeddings. arXiv:1603.09188 [cs]. arXiv:1603.09188

Gildenblat, J. (2023). PyTorch library for CAM methods. GitHub. https://github.com/charlespwd/project-title.

Gray, D., Yu, K., Xu, W., & Gong, Y. (2010). Predicting Facial Beauty without Landmarks. In K. Daniilidis, M. Petros, & N. Paragios (Eds.), Computer Vision – ECCV 2010. Lecture Notes in Computer Science. Berlin, Heidelberg Springer pp. 434–447

Harpaintner, M., Trumpp, N. M., & Kiefer, M. (2018). The Semantic Content of Abstract Concepts: A Property Listing Study of 296 Abstract Words. *Frontiers in Psychology, 9*, 1748. https://doi.org/10.3389/fpsyg.2018.01748

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR). Las Vegas, NV, USA: IEEE pp. 770–778. Available from: http://ieeexplore.ieee.org/document/7780459/.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33*, 6840–6851.

Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., & Agha, Z., et al. (2017). Automatic understanding of image and video advertisements. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 1705–1715.

Ibrahim, R., & Shafiq, M.O. (2023). Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Comput Surv,* 55(10) https://doi.org/10.1145/3563691.

Instagram - Claire Barrow. (2023). https://www.instagram.com/claire_barrow/. Accessed 18 August 2023

Instagram - Mikey Joyce. (2023). https://www.instagram.com/m___joyce/. Accessed 18 August 2023

Jansson, J., & Hracs, B. J. (2018). Conceptualizing curation in the age of abundance: The case of recorded music. *Environment and Planning A: Economy and Space, 50*(8), 1602–1625.

Jeanneret, G., Simon, L., & Jurie, F. (2022). Diffusion Models for Counterfactual Explanations. CoRR. https://doi.org/10.48550/arXiv.2203.15636. arXiv:2203.15636

Joo, J., Li, W., Steen, F.F., & Zhu, S.C. (2014). Visual Persuasion: Inferring Communicative Intents of Images. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 216–223. Available from: https://openaccess.thecvf.com/content%5Fcvpr%5F2014/html/Joo%5FVisual%5FPersuasion%5FInferring%5F2014%5FCVPR%5Fpaper.html.

Jung, H., & Oh, Y. (2021) Towards Better Explanations of Class Activation Mapping. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE pp. 1316–1324. Available from: https://doi.org/10.1109/ICCV48922.2021.00137.

Kalanat, N., & Kovashka, A. (2022). Symbolic image detection using scene and knowledge graphs. arXiv:2206.04863

Kandel, I., & Castelli, M. (2020). How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. *Applied Sciences, 10*(10), 3359.

Kantharaju, R.B., Langlet, C., Barange, M., Clavel, C., & Pelachaud, C. (2020) Multimodal analysis of cohesion in multi-party interactions. In: Proceedings of the twelfth language resources and evaluation conference. pp. 498–507.

Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Trans Scalable Inf Syst, 7*(24), e2. https://doi.org/10.4108/eai.13-7-2018.159623

Kiela, D., & Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha, Qatar: Association for Computational Linguistics pp. 36–45. Available from: https://aclanthology.org/D14-1005.

Kingma, D.P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Y. Bengio, Y. LeCun (Eds.), 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Available from: arXiv:1412.6980

Koolen, M., Van Gorp, J., & Van Ossenbruggen, J. (2019). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities, 34*(2), 368–385.

Kornblith, S., Shlens, J., & Le, Q.V. (2019). Do Better ImageNet Models Transfer Better? In: IEEE Conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE pp. 2661–2671. Available from: http://openaccess.thecvf.com/content%5FCVPR%5F2019/html/Kornblith%5FDo%5FBetter%5FImageNet%5FModels%5FTransfer%5FBetter%5FCVPR%5F2019%5Fpaper.html.

Kousta, S.T., Vigliocco, G., Vinson, D.P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General,* 140(1), 14–34. Place: US Publisher: American Psychological Association. https://doi.org/10.1037/a0021446.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., et al. (2020). The Open Images Dataset V4. *International Journal of Computer Vision, 128*(7), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Lazaridou, A., Pham, N.T., & Baroni, M. (2015). Combining Language and Vision with a Multimodal Skip-gram Model. arXiv:1501.02598 [cs]. arXiv:1501.02598

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. nature, 521*(7553), 436–444.

Lin, B. (2022) Knowledge Management System with NLP-Assisted Annotations: A Brief Survey and Outlook. In G. Drakopoulos, & E. Kafeza (Eds.), Proceedings of the CIKM 2022 Workshops co-located with 31st ACM International Conference on Information and Knowledge Management (CIKM 2022), Atlanta, USA, October 17-21, 2022. vol. 3318 of CEUR Workshop Proceedings. CEUR-WS.org. Available from: https://ceur-ws.org/Vol-3318/short18.pdf.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). *A survey of transformers. AI Open, 3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Mahendran, A., & Vedaldi, A. (2016). Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. *Int J Comput Vis, 120*(3), 233–255. https://doi.org/10.1007/s11263-016-0911-8

Marino, K., Salakhutdinov, R., & Gupta, A. (2017). The More You Know: Using Knowledge Graphs for Image Classification. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017. Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; pp. 20–28. Available from: https://doi.org/10.1109/CVPR.2017.10

Martinez Pandiani, D.S., & Presutti, V. (2023). Seeing the Intangible: A Survey of Computer Vision-Based Approaches for Abstract Concept Detection in Still Images. arXiv preprint arXiv:2308.10562

Miller, G. A. (1998). WordNet: An electronic lexical database. MIT press

Mitchell, W.T. (1995). Picture theory: Essays on verbal and visual representation. University of Chicago Press

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data Soc, 3*(2), 205395171667967.

Mo, Y., Wu, Y., Yang, X., Liu, F., & Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing, 493*, 626–646. https://doi.org/10.1016/j.neucom.2022.01.005

Mohammad, S.M., & Kiritchenko, S. (2018). WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, & K. Hasida, et al. (Eds.), Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA). Available from: http://www.lrec-conf.org/proceedings/lrec2018/summaries/966.html.

Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., & Paulson, P. (2008). The open provenance model: An overview. In: Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers 2. Springer pp. 323–326.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Ng, H., Nguyen, V.D., Vonikakis, V., & Winkler, S. (2015). Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. In Z. Zhang, P. Cohen, D. Bohus, R. Horaud, & H. Meng (Eds.), Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015. ACM pp. 443–449. Available from: https://doi.org/10.1145/2818346.2830593.

Nguyen, A.M., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In DD. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, December 5-10, 2016, Barcelona, Spain pp. 3387–3395. Available from: https://proceedings.neurips.cc/paper/2016/hash/5d79099fcdf499f12b79770834c0164a-Abstract.html.

Nguyen, A., Yosinski, J., & Clune, J. (2019). Understanding Neural Networks via Feature Visualization: A Survey. In W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, & K. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. vol. 11700 of Lecture Notes in Computer Science. Springer pp. 55–76. Available from: https://doi.org/10.1007/978-3-030-28954-6%5F4.

Offert, F. (2019). Images of Image Machines. Visual Interpretability in Computer Vision for Art. In: Computer vision–ECCV 2018 workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II 15. Springer pp. 710–715. Available from: https://openaccess.thecvf.com/content%5Feccv%5F2018%5Fworkshops/w13/html/Offert%5FImages%5Fof%5FImage%5FMachines.%5FVisual%5FInterpretability%5Fin%5FComputer%5FVision%5Ffor%5FECCVW%5F2018%5Fpaper.html.

Offert, F., & Bell, P. (2021). Understanding perceptual bias in machine vision systems. INFORMATIK.

Offert, F., & Bell, P. (2021). Perceptual bias and technical metapictures: critical machine vision as a humanities challenge. *AI & SOCIETY, 36*, 1133–1144.

Ortis, A., Farinella, G.M., & Battiato, S. (2020). Survey on Visual Sentiment Analysis. *IET Image Processing,* 14(8), 1440–1456. ArXiv: 2004.11639. https://doi.org/10.1049/iet-ipr.2019.1270.

Pandiani, D. S. M., & Presutti, V. (2022). Coded Visions: Addressing Cultural Bias in Image Annotation Systems with the Descriptions and Situations Ontology Design Pattern. In: Proceedings of the The 6th international conference on Graphs and Networks in the Humanities p. 8

Panofsky, E., & Drechsel, B. (1955). Meaning in the visual arts. University of Chicago Press Chicago

Presutti, V., Draicchio, F., & Gangemi, A. (2012). Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames. In A. ten Teije, Völker J, Handschuh S, Stuckenschmidt H, d'Aquin M, & A. Nikolov, et al. (Eds.), Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings. vol. 7603 of Lecture Notes in Computer Science. Springer pp. 114–129. Available from: https://doi.org/10.1007/978-3-642-33876-2_5F12.

Rafferty, P., & Hidderley, R. (2017). *Indexing multimedia and creative works: the problems of meaning and interpretation*. London: Routledge.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. CoRR. abs/2204.06125. https://doi.org/10.48550/arXiv.2204.06125. arXiv:2204.06125

Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., et al. (2019). A review on state-of-the-art violence detection techniques. *IEEE Access, 7*, 107560–107575.

Régimbeau, G. (2014). Image source criticism in the age of the digital humanities. Heritage and Digital Humanities: How Should Training Practices Evolve? 4.

Rittel, H. (1967). Wicked problems. *Management Science,* 4(14)

Rodríguez-Ortega, N. (2020). Image processing and computer vision in the field of art history. In: The Routledge companion to digital humanities and art history. New York : Routledge, 2020: Routledge pp. 338–357.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 10684–10695.

Rosenbaum, S. C. (2011). *Curation nation*. McGraw-Hill.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sabatelli, M., Banar, N., Cocriamont, M., Coudyzer, E., Lasaracina, K., & Daelemans, W., et al. (2021). Advances in Digital Music Iconography: Benchmarking the detection of musical instruments in unrestricted, non-photorealistic images from the artistic domain. *Digital Humanities Quarterly,* 15(1)

Segalin, C., Cheng, D. S., & Cristani, M. (2017). Social Profiling through Image Understanding: Personality Inference Using Convolutional Neural Networks. *Computer Vision and Image Understanding., 156*, 34–50. https://doi.org/10.1016/j.cviu.2016.10.013

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., & Chakraborty, T., et al. (2020). SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor! arXiv:2008.03781 [cs]. arXiv:2008.03781

Sharma, S., & Mehra, R. (2018). Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express, 4*(4), 247–254. https://doi.org/10.1016/j.icte.2018.10.007

Shirali, A., & Hardt, M. (2023), What Makes ImageNet Look Unlike LAION. CoRR. https://doi.org/10.48550/arXiv.2306.15769. arXiv:2306.15769

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *J Big Data, 6*, 60. https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society p. 1–14.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034

Smits, T. (2022). The Visual Digital Turn - Computer Vision and the Humanities. video recording. Available from: https://www.youtube.com/@KBR-BEL.

Smits, T. (2023). In K. Lab (Ed.), Can computer vision find illustrations of nineteenth-century railway crashes?). KB Lab. Available from: https://lab.kb.nl/about-us/blog/can-computer-vision-find-illustrations-nineteenth-century-railway-crashes.

Smits, T., & Wevers, M. (2022). The agency of computer vision models as optical instruments. *Vis commun, 21*(2), 329–349.

Solera, F., Calderara, S., & Cucchiara, R. (2017). From Groups to Leaders and Back. In: Group and crowd behavior for computer vision. Elsevier pp. 161–182. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128092767000102.

Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., & Cucchiara, R. (2019) Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Image analysis and processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20. Springer. pp. 729–740.

Stork, D.G. (2009). Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature. In: Computer analysis of images and patterns: 13th International conference, CAIP 2009, Münster, Germany, September 2-4, 2009. Proceedings 13. Springer p. 9–24.

Stork, L., Weber, A., van den Herik, J., Plaat, A., Verbeek, F., & Wolstencroft, K. (2021). Large-scale zero-shot learning in the wild: Classifying zoological illustrations. *Ecological informatics, 62*, 101222.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: 2017 IEEE International conference on computer vision (ICCV). Venice: IEEE pp. 843–852. Available from: http://ieeexplore.ieee.org/document/8237359/.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society pp. 2818–2826. Available from: https://doi.org/10.1109/CVPR.2016.308.

Tan, W.R., Chan, C.S., Aguirre, H.E., & Tanaka, K. (2016). Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In: 2016 IEEE International conference on image processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016. IEEE pp. 3703–3707. Available from: https://doi.org/10.1109/ICIP.2016.7533051.

Vago, N. O. P., Milani, F., Fraternali, P., & da Silva Torres, R. (2021). Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis. *J Imaging, 7*(7), 106. https://doi.org/10.3390/jimaging7070106

van Lange, M. (2022). Emotional Imprints of War: A Computer-assisted Analysis of Emotions in Dutch Parliamentary Debates, 1945-1989. Bielefeld University Press

van Noord, N. (2022). A survey of computational methods for iconic image analysis. Digit scholarsh humanit.

van Noord, N., & Postma, E. O. (2017). Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognit, 61*, 583–592. https://doi.org/10.1016/j.patcog.2016.06.005

Van Zundert, J. J. (2015). Screwmeneutics and hermenumericals: the computationality of hermeneutics. A new companion to digital humanities. pp. 331–347.

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., & Depaepe, F., et al. (2021). Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement. *Mathematics,* 9(3), 287. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/math9030287.

Varghese, E.B., & Thampi, S.M. (2018). A Deep Learning Approach to Predict Crowd Behavior Based on Emotion. In A. Basu, & S. Berretti (Eds.), Smart Multimedia. Lecture Notes in Computer Science. Cham: Springer International Publishing pp. 296–307.

Vassar, B. (2020). The eclectic iconography of hyperpop. The Michigan Daily

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A.N., et al (2017) Attention is All you Need. In I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, & S.V.N. Vishwanathan et al. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017,* December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008. Available from: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Veltmeijer, E.A., Gerritsen, C., & Hindriks, K. (2021). Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing. p. 1–1. Conference Name: IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2021.3065726.

Vilone, G., & Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. CoRR. arXiv:2006.00093

Wevers, M. (2019). Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In: Proceedings of the 1st international workshop on computational approaches to historical language change. Florence, Italy: Association for Computational Linguistics pp. 92–97. Available from: https://aclanthology.org/W19-4712.

Wevers, M., & Smits, T. (2019). The visual digital turn: Using neural networks to study historical images. Digit scholarsh humanit.

Wevers, M., & Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities, 35*(1), 194–207.

Yang, W., Le, H., Savarese, S., & Hoi, S. (2022). OmniXAI: A Library for Explainable AI. arXiv. https://doi.org/10.48550/ARXIV.2206.01612. arXiv:2060.1612

Ye, K., & Kovashka, A. (2018). ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018. vol. 11219 LNCS. Cham, Springer International Publishing. pp. 868–886.

Ye, K., Nazari, N.H., Hahn, J., Hussain, Z., Zhang, M., & Kovashka, A. (2019). Interpreting the Rhetoric of Visual Advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 43(4), 1308–1323 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2019.2947440

Zemni, M., Chen, M., Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2023). OCTET: Object-aware Counterfactual Explanations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15062–15071

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L. (2022) Scaling Vision Transformers. In: IEEE/CVF Conference on computer vision and pattern recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE pp. 1204–1213. Available from: https://doi.org/10.1109/CVPR52688.2022.01179

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2018). From Facial Expression Recognition to Interpersonal Relation Prediction. *International Journal of Computer Vision, 126*(5), 550–569. https://doi.org/10.1007/s11263-017-1055-1

Zhang, D., Cui, M., Yang, Y., Yang, P., Xie, C., Liu, D., et al. (2019). Knowledge Graph-Based Image Classification Refinement. *IEEE Access, 7*, 57678–57690. https://doi.org/10.1109/ACCESS.2019.2912627

Zhao, S., Ding, G., Huang, Q., Chua, T.S., Schuller, B.W., & Keutzer, K. (2018). Affective Image Content Analysis: A Comprehensive Survey. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization. pp. 5534–5541. Available from: https://www.ijcai.org/proceedings/2018/780.

Zhao, S., Huang, Q., Tang, Y., Yao, X., Yang, J., & Ding, G., et al. (2021). Computational Emotion Analysis From Images: Recent Advances and Future Directions. arXiv:2103.10798 [cs]. arXiv:2103.10798

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. In: 2016 IEEE Conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society pp. 2921–2929. Available from: https://doi.org/10.1109/CVPR.2016.319

Zinnen, M., Madhu, P., Bell, P., Maier, A., & Christlein, V. (2023). Transfer Learning for Olfactory Object Detection. arXiv preprint arXiv:2301.09906

Zinnen, M., Madhu, P., Kosti, R., Bell, P., Maier, A., & Christlein, V. (2022). Odor: The icpr2022 odeuropa challenge on olfactory object recognition. In: 2022 26th International conference on pattern recognition (ICPR). IEEE pp. 4989–4994.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Delfina Sol Martinez Pandiani[1] · **Nicolas Lazzari**[2] · **Marieke van Erp**[3] · **Valentina Presutti**[2]

[1] Department of Computer Science and Engineering, University of Bologna, Bologna 40126, BO, Italy

[2] Department of Languages, Literature, and Modern Cultures, University of Bologna, Bologna 40126, BO, Italy

[3] Digital Humanities Lab, KNAW Humanities Cluster, Amsterdam 1012 DK, NH, Netherlands