Neural Computing and Applications (2021) 33:12263–12282
https://doi.org/10.1007/s00521-021-05893-z

REVIEW

Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview

Giovanna Castellano[1] · Gennaro Vessio[1]

Received: 7 December 2020 / Accepted: 27 February 2021 / Published online: 2 April 2021

Abstract
This paper provides an overview of some of the most relevant deep learning approaches to pattern extraction and recognition in visual arts, particularly painting and drawing. Recent advances in deep learning and computer vision, coupled with the growing availability of large digitized visual art collections, have opened new opportunities for computer science researchers to assist the art community with automatic tools to analyse and further understand visual arts. Among other benefits, a deeper understanding of visual arts has the potential to make them more accessible to a wider population, ultimately supporting the spread of culture.

Keywords Digital humanities · Visual arts · Deep learning · Computer vision · Literary review

1 Introduction

In recent years, due to technological improvements and drastic decreases in costs, a large-scale digitization effort has been made, leading to an increasing availability of large digitized visual art collections, e.g. WikiArt. This availability, coupled with the recent advances in deep learning and computer vision, has opened new opportunities for computer science researchers to assist the art community with automatic tools to analyse and further understand visual arts. Among other benefits, a deeper understanding of visual arts has the potential to make them more accessible to a wider population, both in terms of fruition and creation, thus supporting the spread of culture.

The ability to recognize meaningful patterns in visual artworks is intrinsically related to the domain of human perception. Recognizing stylistic and semantic attributes of an artwork, in fact, originates from the composition of the colour, texture and shape features visually perceived by the human eye. In the past, this task has been tackled using hand-crafted features (e.g. [1–4]). However, despite the promising results of feature engineering techniques, early attempts were affected by the difficulty of capturing explicit knowledge about the attributes to be associated with a particular artist or artwork. Such a difficulty arises because this knowledge is typically associated with implicit and subjective expertise human observers may find difficult to verbalize and conceptualize.

Conversely, representation learning approaches, such as those offered by deep learning models, can be the key to success in extracting useful representations from low-level colour and texture features [5–7]. These representations can assist human experts in various art-related tasks, ranging from object detection in paintings to artistic style categorization, useful for example in museum and art gallery websites.

1.1 Motivations

In light of the growing interest in this research domain, this paper aims to provide an overview of some of the most notable works investigating the application of deep learning-based approaches to pattern extraction and recognition in visual artworks. Visual arts are developed primarily for aesthetic purposes, and are mainly concerned with painting, drawing, photography and architecture. In this paper, we focus our attention only to painting and drawing, being two of the most studied visual arts. It is worth noting that

✉ Gennaro Vessio
gennaro.vessio@uniba.it

[1] Department of Computer Science, University of Bari "Aldo Moro", Bari, Italy

this paper is an extension of a brief overview we presented in [8].

This literary review is mainly oriented towards researchers or IT professionals, who may find it exciting to engage in this context, which is very different, for several reasons, from that of traditional photographic and natural scenes. Nevertheless, the paper could also be of interest to humanists, who can discover advances in deep learning and computer vision that can help support their activities. To this end, the paper is intended to provide the reader not only with a state-of-the-art and future perspective on the topic, but also with some guidelines the reader may find useful for entering this line of research.

1.2 Structure of the paper

Being dedicated to a dual audience, the paper is divided into two parts. The first part, which is reported in Sect. 2, describes some available visual art datasets and the main deep learning methods typically used in this context. The second part, which is reported in Sect. 3, discusses the main research trends with reference to what is described in Sect. 2. Finally, Sect. 4 concludes the paper and outlines high-level directions for further research on the topic.

2 Main datasets and deep learning methods

This section reviews some of the most relevant datasets, as well as the basic principles of the main deep learning methods adopted in the context of digitized paintings and drawings.

2.1 Datasets

A schematic description of the most commonly used and relevant datasets is provided in Table 1. WikiArt[1] (formerly known as WikiPaintings) is currently one of the largest online collections of digitized paintings available. It has been a frequent choice for dataset creation in many of the recent studies and has contributed to several art-related research projects. WikiArt integrates a broad set of metadata including style, period, and series. The included artworks span a wide range of periods, with a particular focus on Modern and Contemporary Art. The dataset is constantly growing and includes not only paintings but also sculptures, sketches, posters, and other artworks. At the time of this writing, the WikiArt dataset includes approximately 170,000 artworks attributed to 171 art movements (some examples are shown in Fig. 1). Likewise, Art500k [9] is a large-scale visual art dataset with over 550,000 digitized artworks with rich annotations. In fact, it provides detailed labels not only related to artist and genre but also to event, place and historical figure. All images were mainly scraped from a few websites, including WikiArt itself, and are low resolution copies of the original artworks.

In addition to these projects, some museums begun to make available to developers, researchers and enthusiasts their art collections. For example, the Rijksmuseum of Amsterdam made available (the first API for data collection was launched in 2013) extensive descriptions of more than a half a million historical art objects, hundreds of thousands of object photographs and the complete library catalogue. The dataset was introduced as part of a challenge and consisted of around 100,000 photographic reproductions of the artworks exhibited in the museum. Since then, the digitally available content has been updated. The Rijksmuseum uses controlled vocabularies to unambiguously describe its collection and bibliographic datasets. These thesauri contain information about, for example, people, locations, events and concepts. Currently, the museum is developing technologies to allow users to make optimal use of Linked Open Data. Similarly, on February 2017, the Metropolitan Museum of Art of New York City, colloquially "The MET", made all the images of public-domain works in its collection available under Creative Commons open access license.[2] In particular, the museum made available for download more than 406,000 images of artworks covering more than five thousand years of art from all over the world, from the classic age to contemporary works.

All the datasets mentioned above are mainly designed to perform classification and retrieval tasks. A few datasets, instead, have been enriched with precise information on objects, for the purpose of object recognition and detection. This is the case, for example, of the People-Art dataset, which provides bounding boxes for the single "person" category [10]. The authors claim the reason for only labeling people is that they occur more frequently than any other object class. A similar purpose is pursued by the Behance-Artistic-Media (BAM!) dataset [11], built from Behance, that is a portfolio website for contemporary commercial and professional artists, containing over ten million projects and 65 million images. Artworks on Behance span many fields, such as sculpture, painting, photography, graphic design, graffiti, and advertising. Unlike other datasets, BAM! collects a rich vocabulary of emotion, media, and content attributes. In particular, six content attributes are considered, corresponding to popular PASCAL VOC categories: bicycle, bird, car, cat, dog, and people. More recently, Shen et al. [12] have made publicly

[1] https://www.wikiart.org.

[2] https://www.metmuseum.org.

Table 1 Schematic overview of some of the most frequently used and relevant datasets

Dataset	# artworks	Main task
WikiArt	~ 170,000	Classification and retrieval
Art500k	~ 550,000	Classification and retrieval
Rijksmuseum	~ 650,000	Classification and retrieval
The MET	~ 400,000	Classification and retrieval
People-Art	~ 4500	Object recognition and detection
BAM!	~ 65,000,000	Object recognition and detection
Brueghel	~ 1500	Object recognition and detection
SemArt	~ 20,000	Multi-modal retrieval
Artpedia	~ 3000	Multi-modal retrieval
WikiArt Emotions	~ 4000	Emotion recognition

Fig. 1 Sample digitized artworks from WikiArt

available a new version of the Brueghel dataset[3] with rich annotations. The dataset contains 1587 artworks made in different media (e.g. oil, watercolour, etc.) and on different materials (e.g. paper, panel copper), describing a wide variety of scenes (e.g. landscape, still life, etc.). The authors selected 10 of the most commonly repeated details in the dataset and annotated the bounding box of the duplicated visual patterns. It is worth noting that only the duplicates for each pattern were annotated and not the complete object classes.

To accommodate a multi-modal retrieval task where paintings are retrieved in accordance with an artistic text, and vice versa, a few datasets provide not only metadata attributes but also artistic comments or descriptions, such as those that commonly appear in catalogues or in museum collections. This is the case of SemArt [13] and Artpedia [14]. The main difference between the two datasets is that Artpedia distinguishes *visual* sentences, describing the visual content of the work, from *contextual* sentences, describing the historical context of the work.

A different point of view was taken in the development of the WikiArt Emotions dataset [15], which includes 4105 artworks with annotations for the emotions they evoked in the observer. The artworks were selected from the WikiArt collection for twenty-two categories (Impressionism, Realism, etc.) from four Western styles (Renaissance, Post-Renaissance, Modern and Contemporary Art). Artworks were crowd-sourced annotated for one or more of twenty categories of emotions, including neutrality. In addition to

[3] http://www.janbrueghel.net/.

emotions, the annotations also concern the depiction of a face and how much the observers liked the artworks.

2.2 Deep learning methods

Deep learning refers to a class of machine learning techniques that exploit hierarchical architectures of information processing layers for feature learning and pattern recognition [6]. The main advantage of deep learning models over classic machine learning algorithms is their ability to learn relevant features directly from data. This is desirable, especially in *perceptual* problems, such as those related to aesthetic perception, since mimicking skills that humans feel natural and intuitive have been elusive for machines for a long time.

Indeed, deep learning has a fairly long history, the basic concepts of which originate from artificial neural network research [16]. The neural network paradigm has its roots in such pillars as the works of McCulloch and Pitts [17] and Rosenblatt [18], and was popularized in the 1980s thanks to the rediscovery of the well-known backpropagation learning algorithm [19], which allows a network to update its parameters to learn the solution to a problem based on training data. However, due to the lack of large-scale training data and limited computation power, neural networks went out of fashion in the early 2000s. This was the reason why some seminal papers on convolutional neural networks, e.g. the paper by LeCun et al. [20], and long short-term memory, e.g. the paper by Hochreiter and Schmidhuber [21], remained rather dormant and were only rediscovered later in the last decade. In recent years, the availability of large annotated datasets, such as ImageNet [22], and the development of high performance parallel computing systems, such as GPUs, have fostered a resurgence of neural networks with breakthroughs in historically difficult tasks, notably image classification and natural language processing. In particular, interest in deep neural networks has grown rapidly since the 2012 edition of the ImageNet Large Scale Visual Recognition Challenge, in which AlexNet has far surpassed all previous traditional algorithms [23]. The applications of artificial neural networks today are innumerable and range from healthcare [24] to bioinformatics [25], from biometrics [26] to cybersecurity [27], and so on.

A rich plethora of deep learning techniques has been proposed in the literature. Among these, the most commonly used in the art domain are discussed below.

2.2.1 Convolutional neural networks

Since their appearance, convolutional neural networks (CNNs) have revolutionized image processing and are now almost universally used in computer vision applications [23, 28]. They are much better suited for image data than traditional fully-connected networks, thanks to their ability to retain the spatial input information during the forward propagation. The two main building blocks of CNNs, in fact, which are *convolutional layers* and *pooling layers*, are, respectively, able to detect the presence of features throughout an image and guarantee, to some extent, a translation invariance property. Popular deep CNNs are AlexNet [23], VGG [29] and ResNet [30].

When used in the artistic domain, a CNN can learn to recognize an artist's visually distinctive features by adapting its filters to respond to the presence of these features in a digitized painting. The fully-connected layers typically stacked on top of the convolutional/pooling layers can then be used to translate the presence and intensity of the filter responses into a single confidence score for an artist. A high confidence score is indicative of the presence of a strong response, while a low score indicates that responses are weak or non-existent (a scheme is depicted in Fig. 2). A classic example is PigeoNET, a CNN conceived for an artist attribution task based on artwork training data [31]. When enriched with a feature visualization technique, such a network can show the regions of the input image that have contributed most to the correct artist attribution, especially in case of multiple authorship. More recently, multi-task models have begun to gain popularity, which provide an effective method to solve separate tasks (artist attribution, period estimation, etc.), tackling them simultaneously. Sharing data representation among tasks, in fact, allows the model to exploit the "semantic entanglement" among them to achieve better accuracy [32].

One of the keys to the success of these models is their ability to "transfer" knowledge from one domain to another, provided that the latter is not too dissimilar from the first [33, 34]. Transfer learning is typically done by *fine-tuning* some of the higher layers of a model previously trained for another (more general) task, continuing backpropagation for the specific prediction task. Fine-tuning only higher-level portions of the network is motivated by the observation that the earlier layers of a pre-trained CNN provide generic features (e.g. edges, colour blobs, etc.) that could be useful for many tasks, while later layers are progressively more specific to the details of the images contained in the original dataset. Re-training the network again on a specific dataset slightly adjusts the more abstract representations learned by the network, in order to make them more tailored for the image domain at hand. Since artwork datasets are typically smaller in size that traditional natural image datasets, such as ImageNet, this re-training step is generally required to improve prediction performance.

It is worth noting that, as done in [35], the features provided by the bottleneck layer of a deep pre-trained

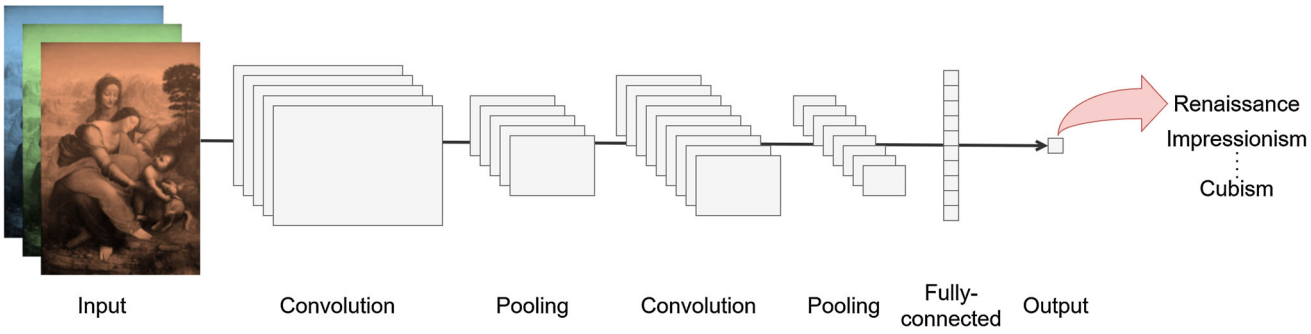

Fig. 2 Schema of a typical CNN architecture. The network is here called to classify the artistic movement to which the painting belongs

model can be used as a *visual embedding* to achieve a more compact feature space representation. Alternatively, compact representations can be learned directly from data using a convolutional autoencoder model [36].

In addition to being used for image classification, CNNs usually form the backbone of many current object detection systems. These involve not only recognizing and classifying objects in an image, but also localizing the position of the objects by drawing a rectangular bounding box around them [37]. This clearly makes object detection more difficult than traditional image classification. Generic object detection frameworks can be mainly categorized into two classes [38]. The first one includes models that initially generate region proposals and then classify each proposal into different categories. This two-step process was pioneered by the well-known R-CNN model [39]. The second class concerns those models that regard object detection as a regression or classification problem, adopting a unified process to obtain categories and locations directly in one step. One of the most popular frameworks that falls into this class is the "You Only Look Once" (YOLO) object detection family [40]. A trade-off should be considered between the two classes, as region proposal-based methods usually perform better, while regression-based methods are faster at the expense of decreased accuracy.

The main issue encountered when using object detectors on artistic images is the so-called *cross-depiction* problem [41, 42], that is the problem of detecting objects regardless of how they are depicted (painted, drawn, photographed, etc.). Most methods tacitly assume photographic input, both at training and test time; however, any solution that does not generalize well regardless of its input depiction is of limited applicability.

A final observation concerns the evaluation of CNN performance. When used for tasks such as style classification or time period estimation, these models are typically evaluated with standard classification and regression metrics, such as accuracy and mean absolute error. On the other hand, in the context of object detection, a lot of attention is paid to metrics such as precision and recall to evaluate the quality of the predicted bounding boxes.

2.2.2 Generative adversarial networks

Generative adversarial networks (GANs), proposed by Goodfellow et al. [43], represent a paradigm for unsupervised deep learning [44]. They are characterized by a pair of networks, typically consisting of convolutional and/or fully-connected layers, which are in competition against each other. The first network, generally referred to as the *generator* $\mathcal{G}$, creates fake images, with the aim of making them as realistic as possible. The second network, called the *discriminator* $\mathcal{D}$, receives both real and fake images, with the aim of telling them apart. The two networks are trained simultaneously. The cost of training is evaluated using a value function V and implies the resolution of $\max_{\mathcal{D}} \min_{\mathcal{G}} V(\mathcal{G}, \mathcal{D})$, where the discriminator tries to maximize its classification accuracy, while the generator tries to deceive the discriminator as much as possible. When the generator is able to perfectly match the real data distribution, then the discriminator is fooled to the maximum, predicting 0.5 for all input images. In other words, $\mathcal{D}$ can no longer distinguish between real samples and fake images.

As generative models learn to capture the statistical distribution of data, this allows for the synthesis of samples from the learned distribution. In the specific context of computational creativity, GANs allow professionals to automatically create a form of art [45]. Unfortunately, GAN training is not easy and often results in the problem of *mode collapse*. This means that the generator always starts exploring the same pattern, producing a small set of very similar samples (i.e. with low diversity) [46].

In this context, it is much more difficult to quantitatively evaluate and compare GAN architectures, as there is no objective loss function used to train the generator and no way to objectively evaluate the progress of training. Usually, practitioners manually assess the quality of the training, generating samples from the generator and

evaluating the plausibility and diversity of the resulting synthetic images.

2.2.3 Recurrent neural networks

Recurrent neural networks (RNNs) are computational learning models with a "memory", meaning that they take their internal parameters not only dependent on the input at current time t, but also on the output at time $t-1$. Thanks to this, they can handle arbitrary input/output sequences, which makes them suitable for temporal and sequential data. Indeed, RNNs are ideal for tasks such as natural language processing [47] and speech recognition [48], in which they have reached state-of-the-art. RNNs have a long history and were already known in the 1980s [49]. The Hopfield network, introduced in 1982 by Hopfield [50], can be considered one of the first networks with recurring connections. Unfortunately, the basic version of RNNs fails to learn long-term dependencies due to the well-known *vanishing gradient* problem. Architectural changes have been proposed to address this problem, making RNNs the powerful tool they are today. These include the aforementioned long short-term memory [21] and gated recurring units [51].

In particular, one of the key findings of RNNs on challenging natural language processing problems is the use of so-called *word embeddings*, which translate large sparse vectors into a smaller space that preserves semantic relationships, thus improving generalization performance. Popular word embeddings are Word2Vec [52] and GloVe [53].

In the particular context of visual arts, RNN models are clearly rarely used alone, as the typical input consists of artistic images, i.e. data having a spatial rather than a temporal nature. However, RNNs are increasingly being used in conjunction with computer vision techniques to solve multi-modal retrieval tasks [13], and have recently been proposed for question answering on art [54] and artwork captioning [55].

Finally, with regard to performance evaluation, as in traditional information retrieval literature approaches that combine convolutional and recurrent neural network models typically rely on metrics based on precision and recall to quantitatively assess the results obtained.

3 Main research trends

Studies involving deep learning approaches for pattern extraction and recognition in paintings and drawings can be broadly classified according to the tasks performed. These tasks have outlined the following main research trends and directions:

- Artwork attribute prediction;
- Information retrieval and artistic influence discovery;
- Object recognition and detection, including near duplicate detection;
- Content generation.

To a lesser extent, the following topics have also been addressed in the literature:

- Artistic to photo-realistic translation;
- Fake detection;
- Representativity;
- Emotion recognition and memorability estimation;
- Visual question answering;
- Artwork captioning.

Figure 3 shows the trend of all these topics in terms of the number of papers published and publication year of the articles reviewed. It can be seen that since 2018 there has been an increasing number of publications on these topics, demonstrating the growing interest of the scientific community in digitized painting and drawing tasks. The following sections are devoted to discussing each topic in detail.

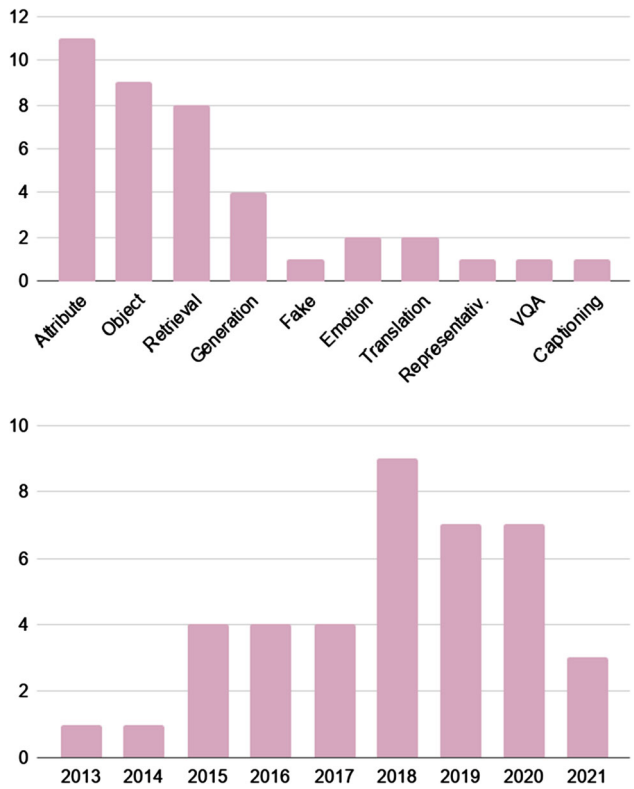

Fig. 3 Paper counting based on topics (above) and year of publication (below)

3.1 Artwork attribute prediction

One of the tasks most frequently faced by researchers in the visual art domain is learning to recognize some artwork attributes (artist, genre, period, etc.) from their *visual style*. Automatic attribute prediction can support art experts in their work on painting analysis and in organizing large collections of paintings. Furthermore, the widespread diffusion of mobile technology has encouraged the tourism industry to develop applications that can automatically recognize the attributes of an artwork in order to provide visitors with relevant information [56, 57].

Although the concept of visual style is rather difficult to define rigorously, distinct styles are recognizable to human observers and are often evident in different painting schools. Artistic visual styles, such as Impressionism, Romanticism, in fact, are characterized by distinctive features that allow artworks to be grouped according to related art movements. In other words, every artwork has a visual style "idiosyncratic signature" [58] which relates it to other similar works.

The papers investigating this topic can be categorized depending on the use of one single model for each individual attribute prediction or a multi-task model aimed at predicting different attributes simultaneously.

3.1.1 Single-task methods

Thanks to their ability to capture not only colour distribution, but also higher level features related to object categories, features automatically extracted by a CNN can easily surpass traditional hand-crafted features when tackling an artwork attribute prediction task. One of the first works on this topic, namely the research presented by Karayev et al. [59], in fact, showed how a CNN pre-trained on PASCAL VOC [60], i.e. an object recognition and detection dataset, is quite effective in attributing the correct painting school to an artwork. The authors explained this behaviour by observing that object recognition depends on the appearance of the object, so the model learns to reuse these features for image style. In other words, they suggest that style is heavily content-dependent.

As mentioned above, another seminal work in this context is the research presented in [31], in which van Noord et al. proposed PigeoNET, a CNN trained on a large collection of paintings to perform the task of automatic artist association based on visual characteristics. These characteristics can also be used to reveal the artist of a precise area of an artwork, in the case of multiple authorship of the same work. We observe that the classification of the unique characteristics of an artist is a complex task, even for an expert. This can be explained by considering that there can be low inter-variability among different artists and high intra-variability in the style of the same artist.

Saleh and Elgammal [61] developed a model capable of predicting not only style, but also genre and artist, based on a metric learning approach. The goal is to learn similarity measures optimized on the historical knowledge available on the specific domain. After learning the metric, the raw visual features are projected into a new optimized feature space on which standard classifiers are trained to solve the corresponding prediction task. In addition to classic visual descriptors, the authors also used features automatically learned by a CNN. Also Tan et al. [62] focused on the three tasks of style, genre and artist classification, and conducted training on each task individually. Interestingly, they also visualized the neurons' responses in the genre classification task, highlighting how neurons in the first layer learn to recognize simple features, while, as layers go deeper, neurons learn to recognize more complex patterns, such as faces in portraits.

Cetinic et al. [63] conducted extensive experimentation to investigate the effective transferability of deep representations across different domains. Interestingly, one of their main findings is that fine-tuning networks pre-trained for scene recognition and sentiment prediction yields better performance in style classification than fine-tuning networks pre-trained for object recognition (typically on ImageNet). A similar investigation was recently conducted by Gonthier et al. [64]. The authors used techniques to visualize the network internal representations to provide clues to understand what a network learns from artistic images. Furthermore, they showed that a double fine-tuning involving a medium-sized artistic dataset can improve classification on smaller datasets, even when the task changes.

Chen et al. [65] further advanced research on the use of CNNs for style classification, moving from the observation that different layers in existing deep learning models have different feature responses for the same input image. To take full advantage of the information from different layers, the authors proposed an adaptive cross-layer model that combines responses from both lower and higher layers to capture style. Finally, another contribution was provided by Sandoval et al. [66], who proposed a two-stage image classification approach to improve style classification. In the first stage, the method splits the input image into patches and uses a CNN model to classify the artistic style for each patch. Then, the probability scores given by the CNN are incorporated into a single feature vector that is provided as an individual input to a shallow neural network model that performs the final classification (see Fig. 4). The main intuition of the proposed method is that individual patches work as independent evaluators for different portions of the

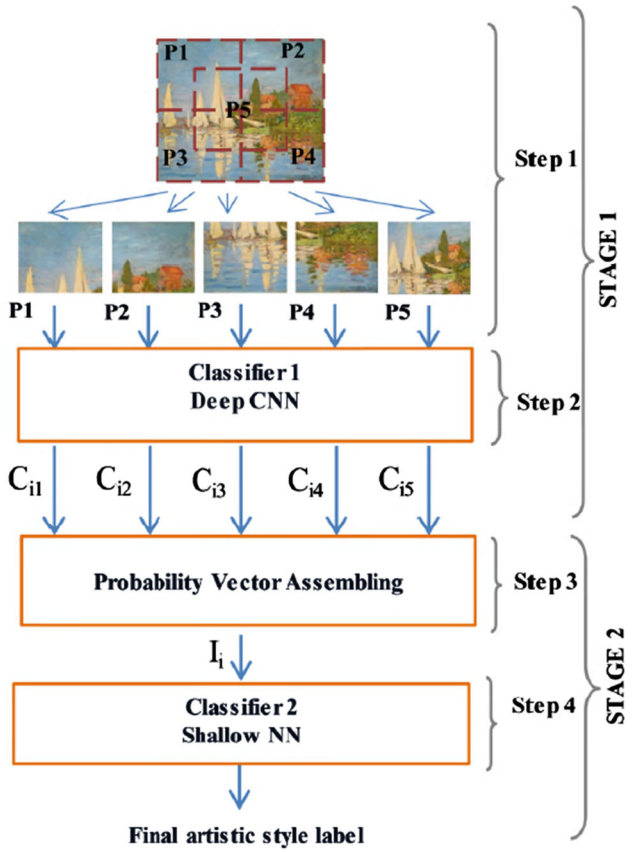

Fig. 4 Two-stage style classification model proposed in [66]. In the first stage, the analyzed images are divided into five patches (P1−P5) and a deep CNN model is used to categorize the style for each patch. In the second stage, the intermediate CNN classification results (probability vectors C1−C5) for the individual patches are assembled into a single input vector fed into a shallow neural network that is trained to provide the final style label

same image; the final model ensembles those evaluations to make the final decision. As is usually the case in this research, confusion was found between historically similar styles. Hence, we conclude that separating visual styles is still a challenging problem.

3.1.2 Multi-task methods

The methods described above address each prediction task individually. Tackling multiple tasks with a single end-to-end trainable model can help in training efficiency and improve classification performance if there is a correlation between different representations of the same input for different tasks. A popular multi-task method is OmniArt [32]. Basically, it consists of a multi-output CNN model in which there is a shared convolutional base for feature extraction and separate output layers, one for each task. The overall training is carried out by minimizing an aggregated loss obtained as a weighted combination of the separate losses.

A different approach was adopted in Belhi et al. [67] who presented a multi-modal architecture that simultaneously takes both digital images and textual metadata as input. The three-channel image is propagated through the convolutional base of a standard ResNet; some metadata, particularly information on genre, medium and style, are one-hot-encoded and provided as input to a shallow feed-forward network. Higher level visual and textual features are concatenated and used to feed the final classification layer. Results indicate that the multi-modal classification system outperforms the individual classification in most cases.

Garcia et al. [35] have gone a step further by combining a multi-output model trained to solve attribute prediction tasks based on visual features and a second model based on non-visual information extracted from artistic metadata encoded using a knowledge graph (see Fig. 5). In short, a knowledge graph is a complex graph that is capable of capturing unstructured relationships between the data represented in the graph. The second model based on the constructed graph is therefore intended to inject "context" information to improve the performance of the first model. To encode the knowledge graph information into a vector representation, the node2vec model [68] was adopted. Indeed, at test time, the context embeddings obtained by computing the knowledge graph cannot be obtained from samples that have not been included as nodes, so the modules that process this information are thrown away. However, the assumption is that the main classification model was forced to learn how to incorporate some contextual information during training. It is worth noting that the proposed method was successfully used by the authors to perform both classification and retrieval.

3.2 Information retrieval and artistic influence discovery

Another task that has attracted attention is finding similarity relationships between artworks of different artists and painting schools. These relationships can help art historians to discover and better understand the influences and changes from an artistic movement to another. Indeed, art experts rarely analyze artworks as isolated creations, but typically study paintings within broad contexts, involving influences and connections among different schools. Traditionally, this kind of analysis is done manually by inspecting large collections of human annotated photos. However, manually searching over thousands of pictures, spanned across different epochs and painting schools, is a very time consuming and expensive process. An automatic support tool would avoid this cumbersome process. More

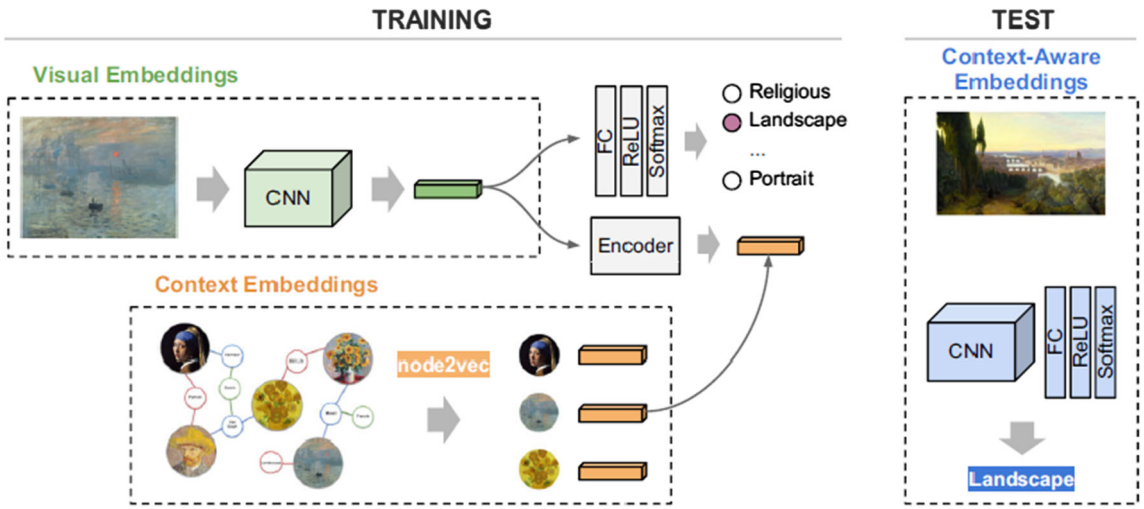

Fig. 5 Scheme of the method proposed in [35] which combines visual and context embeddings. At training time, visual and context embeddings are calculated from the painting image and from the knowledge graph, respectively, and used to optimize the model weights. At test time, to obtain context-aware embeddings from unseen test samples, painting images are fed into a second model

generally, studying how to automatically understand art is a step towards the long-term goal of providing machines with the human aesthetic perception and the ability to semantically interpret images.

This task has been mainly addressed by employing a uni-modal retrieval approach based only on visual features. A different way to look at this problem is to use a multi-modal retrieval approach where computer vision and natural language processing converge towards a unified framework for pattern recognition. These aspects are treated separately in the following subsections.

3.2.1 Uni-modal retrieval

A uni-modal approach to finding similarities among paintings was proposed by Saleh et al. [69], based on traditional hand-crafted features. The authors trained discriminative and generative models for the supervised task of classifying painting style to ascertain which type of features would be most useful in the art domain. Then, once they found the most appropriate features, i.e. those that achieve the highest accuracy, they used these features to judge the similarity between paintings by using distance measures.

A method based on deep learning to retrieve common visual patterns shared among paintings was proposed by Seguin et al. [70]. The authors compared a classic bag-of-words method and a pre-trained CNN in predicting pairs of paintings that an expert considered to be visually related to each other. The authors have shown that the CNN-based method is able to surpass the more classic one. The authors used a supervised approach in which the labels to be predicted were provided manually by human experts.

The manual annotation of images is a slow, error-prone and highly subjective process. Conversely, a completely unsupervised learning method would provide a useful alternative. Gultepe et al. [71], applied an unsupervised feature learning method based on *k*-means to extract features which were then fed into a spectral clustering algorithm for the purpose of grouping paintings. In line with these ideas, in [72, 73] we have proposed a method aimed at finding visual links among paintings in a completely unsupervised way. The method relies solely on visual attributes automatically learned by a deep pre-trained model, so it can be particularly effective when additional information, such as textual metadata, are scarce or unavailable. Furthermore, a computerized suggestion of influences between artists is obtained by exploiting the graph of painters obtained from the visual links retrieved. The analysis of the network structure provides an interesting insight into the influences between artists that can be considered the result of a historical knowledge discovery process (see Fig. 6).

In [36, 74], we have moved further on this direction by exploiting a deep convolutional embedding framework for unsupervised painting clustering, where the task of mapping the raw input data to an abstract, latent space is jointly optimized with the task of finding a set of cluster centroids in the latent feature space. We observed that when the granularity of clustering is coarse, the model takes into account more general features, mainly related to the artistic style. Conversely, when the granularity is finer, the model begins to use content features and tends to group works regardless of the corresponding style. This abstraction capability could be exploited to find similarities between artworks despite the way they are depicted.

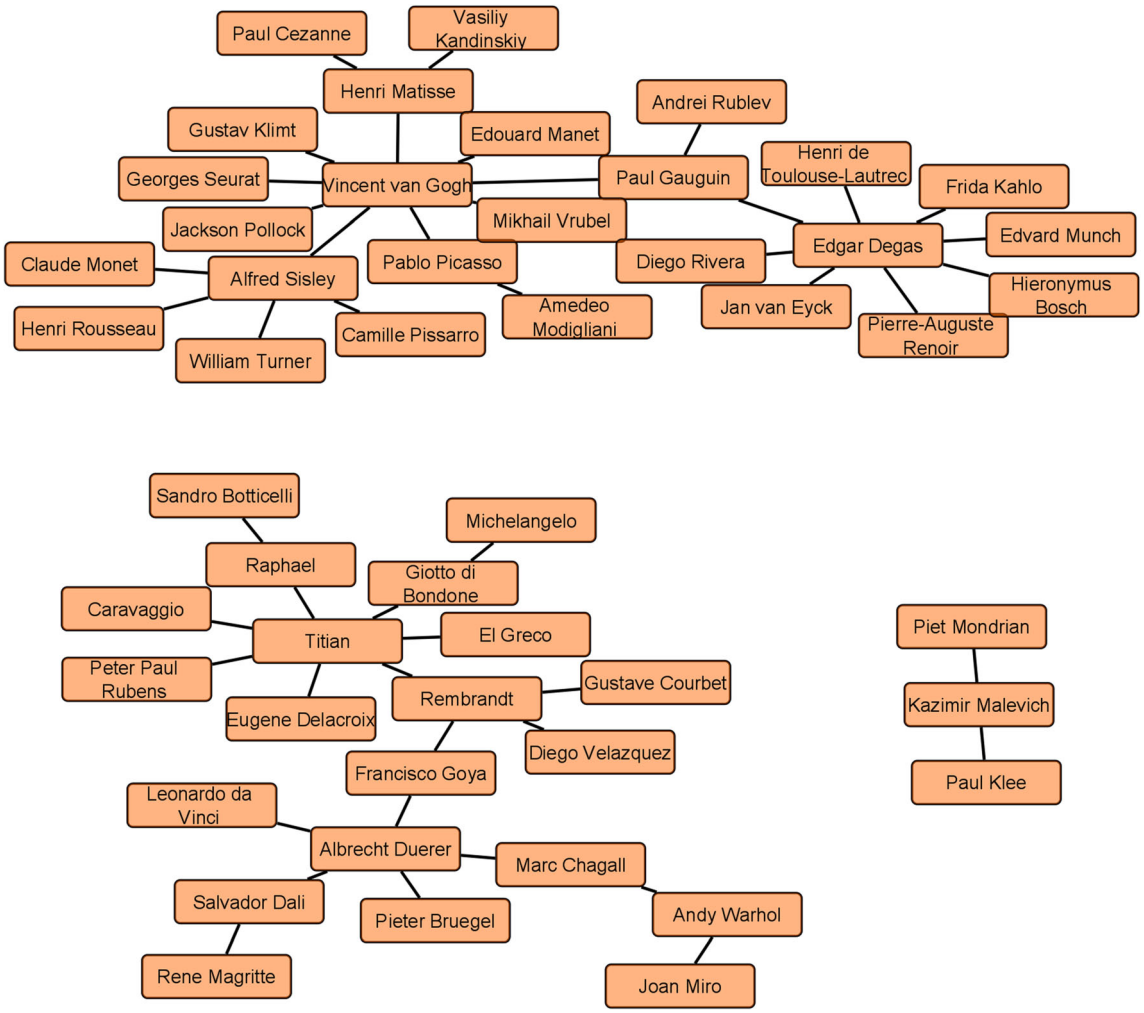

Fig. 6 Influence graph between very famous painters [73]. The connections between painter nodes are established on the basis of the visual similarity between their artworks. Homogeneous groups that share some stylistic characteristics are clearly recognizable

In general, most of the works in the visual art domain adopts a supervised learning approach which, despite accurate results, brings with it the difficulty of having labelled data available. Unsupervised learning has been less studied and we believe it deserves further investigation as a viable alternative to extract useful knowledge from visual data.

3.2.2 Multi-modal retrieval

The first corpus that provides not only painting images and their attributes, but also artistic comments intended to carry out semantic art understanding is the aforementioned SemArt dataset [13]. To evaluate and benchmark this task, Garcia and Vogiatzis designed the Text2Art challenge as a multi-modal retrieval task whose purpose is to assess whether the model is able to match a textual description to the correct painting, and vice-versa. The authors proposed several models that basically share the same working scheme: first, images, descriptions and metadata attributes are encoded into visual and textual embeddings, respectively; then, a multi-modal transformation model is applied to map these embeddings into a common feature space in which a similarity function is used. Although experiments with human evaluators showed that the proposed approaches were unable to achieve art understanding at the human level, the proposed models were able to learn representations that are significant for the task. Indeed, semantic art understanding appears to be a difficult task to solve.

It is worth noting that the same task was pursued by Garcia et al. [35] with the context-based method mentioned above for multi-task attribute prediction.

In a series of papers, Baraldi and colleagues studied methods for aligning text and illustrations in documents, i.e. understanding which parts of a plain text could be related to parts of the illustrations. In [75], in particular, the authors considered the problem of understanding whether a commentary written by an expert on a book, specifically a

digitized version of the Borso d'Este Holy Bible, has some parts that refer to miniature illustrations. To tackle this challenging task, the authors proposed to create a shared embedding space, in which visual and textual features are projected and compared using a distance measure. The so-called BibleVSA dataset was proposed in this study.

In [14], the authors promoted research in this domain by extending the task of visual-semantic retrieval to a setting in which the textual domain does not contain exclusively *visual sentences*, i.e. those that describe the visual content of the work, but also *contextual sentences*, which describe the historical context of the artwork, its author, the place where the painting is located, and so on. To address this two-challenge task, the authors proposed the aforementioned Artpedia dataset. On this dataset, the authors experimented with a multi-modal retrieval model that jointly associates visual and textual elements, and discriminates between visual and contextual sentences of the same image.

Considering that artistic data are often smaller in size than traditional natural image datasets, the authors extended their previous work by moving to a semi-supervised paradigm, in which knowledge learned on ordinary visual-semantic datasets (source domains) is transferred to the artistic (target) domain [76]. As source domains, the authors used Flickr30k [77] and MS COCO [78], which are composed of natural images and are commonly used to train multi-modal retrieval methods. Instead, the aforementioned BibleVSA and SemArt datasets were used as target domains. Experiments validated the proposed approach and highlighted how the distributions of the target and source domains are significantly separated in the embedding space (Fig. 7). This emphasizes that artistic datasets define a completely new domain compared to ordinary datasets.

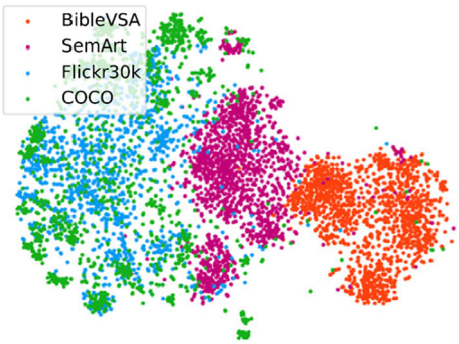

Fig. 7 Comparison of the visual and textual features of ordinary visual-semantic datasets (Flickr30k and MS COCO) and BibleVSA and SemArt [76]. While features from Flickr30k and MS COCO mostly overlap, features from BibleVSA and SemArt appear clearly separated from the other two and from each other. The two-dimensional visualization was achieved by the authors using the well-known t-SNE algorithm [80] on top of the features. These were obtained with a standard VGG architecture; other deep nets were used in the paper, yielding similar results

3.3 Object recognition and detection

Another task often faced by the research community working in this field is finding objects in artworks. Recognizing and detecting objects in artworks can help solve large-scale retrieval tasks as well as support the analyses made by historians. Indeed, art historians are often interested in finding out when a specific object first appeared in a painting or how the representation of an object evolved over time. A pioneering work in this context has been the research reported in [41] by Crowley and Zisserman. They proposed a system that, given an input query, retrieves positive training samples by crawling Google Images on-the-fly. These are then processed by a pre-trained CNN and used together with a pre-computed pool of negative features to learn a real-time classifier. Finally, the classifier returns a ranked list of paintings containing the queried object.

In this context, Cai et al. [42, 79] were among the first to emphasize the importance of addressing the *cross-depiction* problem in computer vision, i.e. recognizing and detecting objects regardless of whether they are photographed, painted, drawn, and so on. The variance between photos and artworks is greater than either domains when considered alone, so classifiers usually trained on natural images may encounter difficulties when it comes to painting images, due to the domain shift. Given the limitless range of potential depictions of the same object, the authors acknowledge that a candidate solution is not learning the specificity of each representation, but learning the abstraction that different representations share so that they can be recognized independently of their representation.

Crowley and Zisserman [81] have improved their previous work by moving from image-level classifiers, i.e. those that take the overall image as input, to object detection systems, which are assumed to improve detection of small objects which are particularly prevalent in paintings. The results of their experimental study have provided evidence that detectors can find many objects in paintings that would likely have been overlooked by object recognition methods. Westlake et al. [10] showed how a CNN only trained on photos can lead to overfitting; on the contrary, fine-tuning on artworks allows the model to generalize better to other artwork styles. To evaluate their proposal, the authors used the People-Art dataset, purposely realized for the task of detecting people. To push forward research in this direction, Wilber et al. [11] did not focus on people, but proposed to use the aforementioned BAM! dataset, designed to provide researchers with a large

benchmark dataset to expand the current state-of-the-art of computer vision to the visual art domain.

More recently, Gonthier et al. [82] focused on more specific objects or visual attributes that may be useful to art historians, such as ruins or nudity, and iconographic characters, such as the Virgin, Jesus. These categories are unlikely to be inherited directly from photographic databases. To overcome this problem, the authors proposed a "weakly supervised" approach that can learn to detect objects based only on image-level annotations. The goal is to detect new, unseen objects with minimal supervision.

A different perspective from which object detection can be viewed in this context is *near duplicate detection*. This task is not to find distinct instances of a same object class, but to automatically discover nearly identical patterns in different images. In [12], Shen et al. addressed this problem by applying a deep neural network model to a dataset of artworks attributed to Jan Brueghel purposely annotated by the authors. The key technical insight of the method is to adapt a deep standard feature to this task, perfecting it on the specific art collection using self-supervised learning. Spatial consistency between adjacent feature matches is used as a supervisory fine-tuning signal. The fitted function leads to a more accurate style invariant match and can be used with a standard discovery approach, based on geometric verification, to identify duplicate patterns in the dataset. The method is self-supervised, which means that the training labels are derived from the input data. Ufer et al. [83] recently extended this research by presenting a multi-style fusion approach that successfully reduces the domain gap and improves retrieval results in larger collections with a large number of distractors.

As can be realistically supposed, the approaches discussed in this section are more successful with artworks that are "photo-realistic" by nature, but can fail or show degraded performance when used on more abstract styles such as Cubism or Expressionism. Such abstract styles pose serious challenges since the depictions of objects and subjects may show strong individualities and are therefore difficult to represent through generalizable patterns.

3.4 Content generation

A central problem in the Artificial Intelligence community is generating art through machines: in fact, making a machine capable of showing creativity on a human level (not only in painting, but also in poetry, music, and so on) is widely recognized as an expression of intelligence. Traditional literature on computational creativity has developed systems for art generation based on the involvement of human artists in the generation process (see, for example, [84]). More recently, the advent of the Generative Adversarial Network paradigm has allowed researchers to develop systems that do not put humans in the loop but make use of previous human products in the learning process. This is consistent with the assumption that even human experts use prior experience and their knowledge of past art to develop their own creativity.

Elgammal et al. [45], proposed CAN (Creative Adversarial Network): a variant of a classic GAN architecture that aims to create art by maximizing deviation from established styles and minimizing deviation from art distribution. In other words, the model "tries to generate art that is novel, but not too novel". Deviating from established styles is important, as a classic GAN would "emulate" previous data distribution showing limited creativity. The effectiveness of CAN was assessed by involving the judgment of human evaluators, who regularly confused generated art with human art. Examples of images generated by CAN are shown in Fig. 8. Of course, the machine does not have a semantic understanding of the subject, since, as mentioned above, its learning is based only on exposure to the prior art.

Similarly, Tan et al. [85, 86] proposed ArtGAN: a GAN variant in which the gradient information with respect to the label is propagated from the discriminator back to the generator for better learning representation and image quality. Further architectural innovations are introduced in the model generation. Qualitative results show that ArtGAN is capable of generating plausible-looking artworks based on style, artist and genre.

Lin et al. [87] observed that attributes such as artist, genre and period can be crucial as control conditions to guide the painting generation. To this end, they proposed a multi-attribute guided painting generation framework in which an asymmetrical cycle structure with control branches is used to increase controllability during content generation.

Generating content with GANs is a field that has received tremendous interest in recent times. The interested reader can refer to the inspiring paper about style-based generator [88] and GAN applications quite related to art like anime [89] and fashion [90] design.

3.5 Other topics

There are a number of topics that have been studied less thoroughly by researchers but which deserve to be mentioned. They are briefly described below.

3.5.1 Artistic to photo-realistic translation

Tomei at al. [91, 92] observed that the poor performance usually provided by state-of-the-art deep learning models is due to the natural images they are pre-trained on: this results in a gap between high-level convolutional features

Fig. 8 Sample images generated by CAN, ranked as highly plausible by human experts [45]

of the natural and artistic domain, whose visual appearance differ greatly. To reduce the shift between feature distributions from the two domains, without the need to re-train the deep models, the authors proposed Art2Real: instead of fine-tuning previous models on the new artistic domain, the method translates artworks directly in photo-realistic images. The model generates natural images by retrieving and learning details from real photos through a similarity matching strategy that leverages a weakly supervised understanding of the scene. Experimental results show that the proposed technique leads to increased realism and reduces the domain shift.

Another way to perform image-to-image translation, particularly in the opposite direction, is to use the well-known *neural style transfer* technique originally proposed by Gatys et al. [93]. This consists in combining the content of one image with the style of another, typically a painting, to synthesize a combined unedited image. Unfortunately, while effective in transferring artistic styles, this method usually works poorly in the opposite direction, i.e. when asked to translate artworks into photo-realistic images. Since there are many studies exploring how to automatically transform photo-realistic images into synthetic artworks, and some literary reviews, such as [94], have already been done, the topic of neural style transfer is not covered in this paper.

3.5.2 Fake detection

An essential task for art experts is to judge the authenticity of an artwork. Historically, this task was based on the search for detailed "invariant" characteristics of an artist's style regardless of composition or subject matter. Currently, the analysis of these details is supported by techniques based, among others, on chemical and radiometric analyses. State-of-the-art computer vision techniques have to potential to provide cost-effective alternatives to the sophisticated analyses performed in laboratory settings. To this end, Elgammal et al. [95] proposed a computerized approach to analyzing strokes in artists' line drawings to facilitate the attribution of drawings without being fooled by counterfeit art. The proposed methodology is based on the quantification of the characteristics of individual strokes by combining different hand-crafted and learned features. The authors experimented with a collection of drawings mainly by Pablo Picasso, Henry Matisse, and Egon Schiele, showing good performance and robustness to falsely claimed works.

3.5.3 Representativity analysis

In a very recent work, Deng et al. [96] have proposed the concept of *representativity* to quantitatively assess the extent to which a given individual painting can represent the general characteristics of an artist's creation. To tackle this task, the authors proposed a novel deep representation of artworks enhanced by style information obtained

through a weighted pooling feature fusion module. Then, a graph-based learning method is proposed for representativity learning, which considers intra-category and extra-category information. Since historical factors are significant in the art domain, the time of creation of a painting is introduced into the learning process. User studies showed that the proposed approach helps to effectively access artists' creation characteristics by ordering paintings in accordance with representativity from highest to lowest.

3.5.4 Emotion recognition and memorability estimation

With the rise of visual data online, understanding the feelings aroused in the observer by visual content is gaining more and more attention in research. However, there are few works in the literature that address this challenging task, mainly due to the inherent complexity of the problem and the scarcity of digitized artworks annotated with emotional labels.

Lu et al. [97] proposed an adaptive learning approach to understand the emotional appeal of paintings. The method uses labelled photographs and unlabelled paintings to distinguish positive from negative emotions and to differentiate reactive emotions from non-reactive ones. The learned knowledge of photographs is transferred to paintings by iteratively adapting feature distribution and maximizing the joint likelihood of labelled and unlabelled data.

Cetinic et al. [98, 99], investigated the possibility of using learned visual features to estimate the emotions evoked by art as well as painting *memorability*, i.e. how easy it is for a person to remember an image. In fact, people have been shown to share a tendency to remember the same images, indicating that memorability is universal in nature and lies beyond our subjective experience [100]. This also indicates that some image features contribute more to memorability than others. The authors used a model trained to predict memorability scores of natural images to explore the memorability of artworks belonging to different genres and styles. Experiments showed that nude and portrait paintings are the most memorable genres, while landscape and marine paintings are the least memorable. As for image style, it turned out that abstract art is more memorable than figurative art. Additionally, an analysis of the correlation between memorability and image features related to composition, colour and the visual sentiment response evoked by abstract images was provided. Results showed that there is no correlation between symmetry and memorability, however memorability is positively correlated with the likelihood of an image to evoke a positive feeling. Their results also suggest that content and image lighting have a significant influence on aesthetics, in which colour vividness and harmony strongly influence the prediction of sentiment, while the emphasis on objects has a strong impact on memorability.

3.5.5 Visual question answering

Recently, Garcia at al. [54] have built, on top of the previously proposed SemArt, the AQUA dataset, which aims to be a preliminary benchmark for *visual question answering* in the art domain. This refers to the task of providing the computer vision system with a text-based question and the system should give to the user an answer. Baseline results have been presented by the authors with a two-branch model, in which visual and knowledge questions are handled independently.

3.5.6 Artwork captioning

Cetinic recently noted, in [55], that while *image captioning* has been extensively studied in recent literature, little work has been done in the art domain. Image captioning refers to the task of recognizing objects and their relationships in an image to generate syntactically and semantically correct textual descriptions of the image. She conducted an experiment with state-of-the-art methods finding out that it is possible to generate meaningful captions from art, which show strong relevance to the historical-artistic context of the image.

3.6 Summary

A schematic overview of the studies reviewed in this paper is provided in Table 2. The studies are ordered in chronological order to provide a final historical perspective on the research topic and, for each of them, the main task, the method used and the results obtained are summarized.

4 Concluding remarks and future directions

The growing availability of large digitized artwork collections has given rise to a new, intriguing area of research where computer vision and visual arts meet. The new research area is framed as a sub-field of the constantly growing *digital humanities*, which aims to bring together digital technologies and humanities. Applications are innumerable and range from information retrieval in digital databases to the synthetic generation of some form of art.

It is worth pointing out that there are at least two (implicit) assumptions made by researchers in this field. First, the input to deep learning models is assumed to be faithful photographic reproductions of actual paintings. While this can be generally considered true, it is worth remembering that reproduction is highly dependent on the quality of

Table 2 Summary of the studies reviewed. (Note that in the case of partially overlapping works by the same authors, only one article is reported here)

Reference	Main task	Method	Main finding
Karayev et al. [59]	Style recognition	CNN pre-trained on general object recognition data	Transfer learning from traditional photographic domains is effective for visual style classification
Crowley and Zisserman [41]	Object recognition and detection	CNN object classifiers learned from Google Images	Many objects prove elusive, particularly when there are large differences between the representation of the object in natural images and in paintings
Cai et al. [79]	Object recognition and detection	CNN-based features fed into SVM	The distributions of photographic and artistic domain features differ more strongly than those of conventional domain adaptation research
Saleh and Elgammal [61]	Style, genre and artist recognition	Similarity learning based on visual features	A machine is able to express semantic judgments related to aesthetics
van Noord et al. [31]	Author attribution	CNN with an occlusion sensitivity test method to obtain visualizations for artists' characteristic regions	A CNN can generate a view that indicates for each location of an artwork who is the artist who most likely contributed to the visual characteristics of that location
Crowley and Zisserman [81]	Object recognition and detection	Combination of detection and image-level classification	Classifiers learned for regions rather than the whole image recognizes and locates a wide range of small objects in paintings that are not detected by image-level classifiers
Seguin et al. [70]	Visual link retrieval	Search in the embedded space provided by a CNN	Pre-trained CNNs may work better for retrieving visual links than other computer vision methods aimed at analyzing photographs
Tan et al. [62]	Style, genre and artist recognition	CNN with visualizations of the filters learned	For more structured paintings, the visualizations show that CNNs are able to find key objects to classify them
Westlake et al. [10]	Object recognition and detection	Fast R-CNN	A CNN trained on photos only overfits photos, while fine-tuning to artworks allows the model to generalize better to art styles
Elgammal et al. [45]	Painting generation	CAN	Art can be generated by maximizing the deviation from established styles and minimizing the deviation from art distribution
Strezoski and Worring [32]	Author, material and type prediction, and period estimation	Multi-output CNN	Creating a shared representation between tasks allows the model to take advantage of semantic entanglement between them for better performance
Wilber et al. [11]	Object recognition and detection	Comparison of object detection and recognition models	The Behance Artistic Media dataset is proposed with baseline results
Baraldi et al. [75]	Text and document illustration alignment	Visual and textual representations are encoded in a common space	There is a noticeable shift in domain between ordinary visual semantic datasets and artistic ones
Belhi et al. [67]	Author attribution	Multi-input model based on visual and textual feature learning	The multi-modal approach outperforms the uni-modal one in most cases
Cetinic et al. [63]	Style, genre and artist recognition	Fine-tuning CNN models pre-trained on different domains	Pre-trained model initialization influences fine-tuning performance
Elgammal et al. [95]	Fake detection	Segmentation method to quantify the characteristics of individual strokes in drawings	The invariant characteristics of an artist can be detected with high precision at the stroke level
Garcia and Vogiatzis [13]	Multi-modal retrieval	Model based on visual and textual embeddings	Paintings can be retrieved according to an artistic text and vice versa
Gonthier et al. [82]	Object recognition and detection	Weakly supervised object detection model	Iconographic elements that are very specific to art history cannot be easily learned from natural images
Gultepe et al. [71]	Artwork grouping	Unsupervised feature learning based on *k*-means	Distinctive features can be extracted even without prior knowledge
Tan et al. [86]	Painting generation	ArtGAN	With the feedback of the information on the labels, the generator is able to learn more efficiently and generate images with better quality

Table 2 (continued)

Reference	Main task	Method	Main finding
Cetinic et al. [99]	Emotion recognition	Various hybrid architectures based on convolutional and LSTM components	The emotion evoked in the observer and the memorability of an artwork can be estimated with visual features
Chen et al. [65]	Style recognition	CNN with adaptive cross-layer correlation	Learning the correlations between features of different layers can be more effective for style classification
Sandoval et al. [66]	Style recognition	Two-stage CNN-based classification	The machine is confused by historically similar styles
Shen et al. [12]	Near duplicate detection	Semi-supervised spatially consistent feature learning	Duplicated visual patterns in art collections can be automatically discovered
Stefanini et al. [14]	Multi-modal retrieval	Visual semantic model based on triplet loss	The task of matching paintings and sentences is enriched by the task of identifying which sentences actually describe the visual content of a given image
Tomei et al. [91]	Image-to-image translation	Art2Real	The gap between the distributions of visual features from artistic and realistic images can be reduced by translating paintings into photo-realistic images
Castellano et al. [73]	Visual link retrieval and knowledge discovery	Unsupervised search in the embedded space provided by a CNN combined with a complex network approach	A graph showing influences between artists can be constructed based on visual links between digitized artworks
Cornia et al. [76]	Multi-modal retrieval	Semi-supervised joint visual-semantic embedding model with domain transfer	A strong domain shift between natural images and artworks is observed
Deng et al. [96]	Representativity assessment	Framework that embeds painting styles and author information	The extent to which a painting can represent the characteristics of an artist's creations can be evaluated quantitatively
Garcia et al. [35]	Author, school and type prediction, and period estimation	Multi-task and knowledge graph-based models	Injecting contextual information into the learning process can improve performance
Garcia et al. [54]	Visual question answering	Two-branch model that concatenates both image and textual encoding	First baseline on visual question answering on art
Li et al. [87]	Painting generation	Multi-attribute guided GAN	Higher quality results can be obtained by using multiple attributes as control conditions
Ufer et al. [83]	Near duplicate detection	Method based on multi-style feature fusion and iterative voting	Specific motifs can be successfully found even among several distractors
Castellano and Vessio [74]	Artwork grouping	Deep convolutional embedding clustering	A deep learning framework can be used to group paintings at different levels of granularity
Cetinic [55]	Artwork captioning	Transformer based vision-language pre-trained model	Meaningful captions can be generated that show relevance to the historical context
Gonthier et al. [64]	Style classification	Comparison of pre-trained CNN models	Double fine-tuning can improve classification performance

illumination, the relative distance from the artwork when the photo is taken, etc., which can hinder the following feature extraction process if standard criteria are not followed during digitization. Therefore, caution should be used when using digitized images, especially considering that the problems mentioned above can be exacerbated by the reduction in size and normalization process that images usually undergo before being fed to Convolutional Neural Network models. The second assumption is that digitized artworks will be treated in the same way as traditional digital images when provided as input to Convolutional Neural Networks when used as feature extractors. Even this assumption can be trusted, especially if CNN models are asked to find both low-level and high-level visual features to perform tasks as diverse as style, genre, and artist recognition. However, when these models are asked to recognize or even detect objects regardless of how they are depicted, performance drops significantly.

Throughout the paper we have highlighted the main current research directions in the field of pattern extraction and recognition in paintings and drawings, pointing out the

related open problems. To complement this overview, we outline some high-level future directions here.

Domain generalization A crucial problem with using deep neural nets, as mentioned earlier, is that they are particularly data hungry, so they usually need large-sized training data with difficult to collect category labels. Unfortunately, there are no large annotated data sets in the art domain. Since current computer vision modules are generally based on knowledge pre-trained on natural image datasets, the learned models are biased towards them. This bias can result in poor performance, as the visual appearance of artworks is significantly different from that of photo-realistic images, due to the presence of brush strokes, the specific style of the artist, etc. Approaches dedicated to reduce this domain gap, such as those based on *domain generalization*, which attempt to alleviate this issue by producing models that by design generalize well to novel test domains. represent a remarkable line of research (e.g. [101, 102]).

Neuro-symbolic learning As is usually the case in deep learning applications, the real success of these models is that they automatically project raw pixel values into meaningful embeddings where patterns of interest begin to emerge. However, although this representation capability can be enhanced with visualization techniques, such as saliency [103] and attention [104] maps, which provide a means of highlighting the neurons' response to certain input features, explaining the logic behind what the algorithm discovers, as also pointed out in [105], is still very difficult. Very recent works are investigating *neuro-symbolic* approaches, in which knowledge bases are used to guide the learning process of deep neural networks [106, 107]. This "hybrid" approach has the potential to bridge the gap between visual perception and reasoning, and may play a role in enabling a machine to mimic the complex human aesthetic perception, the underlying processes of which are still largely unknown. Filling this gap can foster the dialogue between computer vision enthusiasts and humanists that currently seems to lack [108].

Social robotics As applications of computer vision algorithms to artistic tasks become more mature, an interesting deployment of these techniques in real-world cases is to incorporate them into social robots. These represent an emerging research field focused on developing a "social intelligence" that aims to maintain the illusion of dealing with a human being. A social robot can be equipped with computer vision modules to provide personalized and engaging museum visit experiences [109, 110].

The use of deep learning in the visual arts is currently an area of great research interest. Encouraged by the growing literature that has recently emerged on the topic, we are confident that this exciting field of research will be strengthened in the future, taking advantage of the rapid advances in deep learning approaches. We believe these approaches will continue to evolve rapidly, thus paving the way for the realization of amazing scenarios in which computer systems will be able to analyze and understand fine arts autonomously.

Acknowledgements Gennaro Vessio acknowledges the financial support of the Italian Ministry of University and Research through the PON AIM 1852414 project.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

Declarations

Conflicts of interest The authors declare they have no conflict of interest.

References

1. Carneiro G, da Silva NP, Del Bue A, Costeira JP (2012) Artistic image classification: an analysis on the printart database. In: European Conference on Computer Vision. Springer, pp 143–157
2. Khan FS, Beigpour S, Van de Weijer J, Felsberg M (2014) Painting-91: a large scale database for computational painting categorization. Mach Vis Appl 25(6):1385–1397
3. Shamir L, Macura T, Orlov N, Eckley DM, Goldberg IG (2010) Impressionism, expressionism, surrealism: automated recognition of painters and schools of art. ACM Trans Appl Percept (TAP) 7(2):8
4. Arora RS, Elgammal A (2012) Towards automated classification of fine-art painting style: a comparative study. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp 3541–3544
5. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
6. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
7. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26
8. Castellano G, Vessio G (2021) A brief overview of deep learning approaches to pattern extraction and recognition in paintings and drawings. In: Pattern recognition. ICPR

International workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III, Springer International Publishing, pp 487–501
9. Mao H, Cheung M, She J (2017) Deepart: learning joint representations of visual arts. In: Proceedings of the 25th ACM International Conference on Multimedia. ACM, pp 1183−1191
10. Westlake N, Cai H, Hall P (2016) Detecting people in artwork with CNNs. In: European Conference on Computer Vision. Springer, pp 825−841
11. Wilber MJ, Fang C, Jin H, Hertzmann A, Collomosse J, Belongie S (2017) BAM! The Behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1202−1211
12. Shen X, Efros AA, Aubry M (2019) Discovering visual patterns in art collections with spatially-consistent feature learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9278–9287
13. Garcia N, Vogiatzis G (2018) How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV)
14. 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: International conference on image analysis and processing. Springer, pp 729−740
15. Mohammad S, Kiritchenko S (2018) Wikiart emotions: an annotated dataset of emotions evoked by art. In: Proceedings of the eleventh international conference on Language Resources and Evaluation (LREC 2018)
16. Zhang GP (2000) Neural networks for classification: a survey. IEEE Trans Syst Man Cybern Part C (Appl Rev) 30(4):451−462
17. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115−133
18. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386−408
19. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533−536
20. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541−551
21. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735−1780
22. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211−252
23. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097−1105
24. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2018) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 19(6):1236−1246
25. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851−869
26. Sundararajan K, Woodard DL (2018) Deep learning for biometrics: a survey. ACM Comput Surv (CSUR) 51(3):1−34
27. Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, Gao M, Hou H, Wang C (2018) Machine learning and deep learning methods for cybersecurity. IEEE Access 6:35365−35381
28. LeCun Y, Bengio Y et al (1995) Convolutional networks for images, speech, and time series. Handb Brain Theory Neural Netw 3361(10):1995
29. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770−778
31. Van Noord N, Hendriks E, Postma E (2015) Toward discovery of the artist's style: learning to recognize artists by their artworks. IEEE Signal Process Mag 32(4):46−54
32. Strezoski G, Worring M (2017) OmniArt: multi-task deep learning for artistic data analysis. arXiv preprint arXiv:1708.00684
33. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks?. In: Advances in neural information processing systems, pp 3320−3328
34. Budnik M, Gutierrez-Gomez E-L, Safadi B, Pellerin D, Quénot G (2017) Learned features versus engineered features for multimedia indexing. Multimed Tools Appl 76(9):11941−11958
35. Garcia N, Renoust B, Nakashima Y (2020) ContextNet: representation and exploration for painting classification and retrieval in context. Int J Multimed Inf Retrieval 9(1):17−30
36. Castellano G, Vessio G (2020) Deep convolutional embedding for painting clustering: case study on Picasso's artworks. In: International conference on discovery science. Springer, pp 68−78
37. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep learning for generic object detection: a survey. Int J Comput Vis 128(2):261−318
38. Zou Z, Shi Z, Guo Y, Ye J (2019) Object detection in 20 years: a survey. arXiv preprint arXiv:1905.05055
39. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 580−587
40. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767
41. Crowley EJ, Zisserman A (2014) In search of art. In: European Conference on Computer Vision. Springer, pp 54−70
42. Cai H, Wu Q, Corradi T, Hall P (2015) The cross-depiction problem: computer vision algorithms for recognising objects in artwork and in photographs. arXiv preprint arXiv:1505.00110
43. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672−2680
44. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y (2019) Recent progress on generative adversarial networks (GANs): a survey. IEEE Access 7:36322−36333
45. Elgammal A, Liu B, Elhoseiny M, Mazzone M (2017) CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms. arXiv preprint arXiv:1706.07068
46. Wiatrak M, Albrecht SV (2019) Stabilizing generative adversarial network training: a survey. arXiv preprint arXiv:1910.00927
47. Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Hum Lang Technol 10(1):1−309
48. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 6645−6649
49. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31(7):1235−1270

50. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554−2558
51. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
52. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1903.026780
53. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532−1543
54. Garcia N, Ye C, Liu Z, Hu Q, Otani M, Chu C, Nakashima Y, Mitamura T (2020) A dataset and baselines for visual question answering on art. In: European conference on computer vision. Springer, pp 92–108
55. Cetinic E (2021) Iconographic image captioning for artworks. In: Del Bimbo A et al (eds) Pattern recognition. ICPR international workshops and challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12663. Springer, Cham
56. Ragusa F, Furnari A, Battiato S, Signorello G, Farinella GM (2020) EGO-CH: dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. Pattern Recognit Lett 131:150−157
57. Torres-Ruiz M, Mata F, Zagal R, Guzmán G, Quintero R, Moreno-Ibarra M (2020) A recommender system to generate museum itineraries applying augmented reality and social-sensor mining techniques. Virtual Reality 24(1):175−189
58. Bar Y, Levy N, Wolf L (2014) Classification of artistic styles using binarized features derived from a deep neural network. In: European Conference on Computer Vision. Springer, pp 71−84
59. Karayev S, Trentacoste M, Han H, Agarwala A, Darrell T, Hertzmann A, Winnemoeller H (2013) Recognizing image style. arXiv preprint arXiv:1903.026783
60. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The PASCAL visual object classes (VOC) challenge. Int J Comput Vis 88(2):303−338
61. Saleh B, Elgammal A (2015) Large-scale classification of fine-art paintings: learning the right metric on the right feature. arXiv preprint arXiv:1505.00855
62. Tan WR, Chan CS, Aguirre HE, 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). IEEE, pp 3703−3707
63. Cetinic E, Lipic T, Grgic S (2018) Fine-tuning convolutional neural networks for fine art classification. Expert Syst Appl 114:107−118
64. Gonthier N, Gousseau Y, Ladjal S (2021) An analysis of the transfer learning of convolutional neural networks for artistic images. In: Del Bimbo A et al. (eds) Pattern recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12663. Springer, Cham
65. Chen L, Yang J (2019) Recognizing the style of visual arts via adaptive cross-layer correlation. In: Proceedings of the 27th ACM international conference on multimedia, pp 2459−2467
66. Sandoval C, Pirogova E, Lech M (2019) Two-stage deep learning approach to the classification of fine-art paintings. IEEE Access 7:41770−41781
67. Belhi A, Bouras A, Foufou S (2018) Leveraging known data for missing label prediction in cultural heritage context. Appl Sci 8(10):1768
68. Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp 855−864
69. Saleh B, Abe K, Arora RS, Elgammal A (2016) Toward automated discovery of artistic influence. Multimed Tools Appl 75(7):3565−3591
70. Seguin B, Striolo C, Kaplan F et al (2016) Visual link retrieval in a database of paintings. In: European Conference on Computer Vision. Springer, pp 753−767
71. Gultepe E, Conturo TE, Makrehchi M (2018) Predicting and grouping digitized paintings by style using unsupervised feature learning. J Cultural Heritage 31:13−23
72. Castellano G, Vessio G (2020) Towards a tool for visual link retrieval and knowledge discovery in painting datasets. In: Italian Research Conference on Digital Libraries. Springer, pp 105−110
73. Castellano G, Lella E, Vessio G (2020) Visual link retrieval and knowledge discovery in painting datasets. Multimed Tools Appl 80:6599−6616
74. Castellano G, Vessio G (2020) Deep convolutional embedding for digitized painting clustering. In: International Conference on Pattern Recognition (ICPR2020). IEEE
75. Baraldi L, Cornia M, Grana C, Cucchiara R (2018) Aligning text and document illustrations: towards visually explainable digital humanities. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, pp 1097−1102
76. Cornia M, Stefanini M, Baraldi L, Corsini M, Cucchiara R (2020) Explaining digital humanities by aligning images and textual descriptions. Pattern Recognit Lett 129:166−172
77. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641−2649
78. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár, Zitnick (2014) Microsoft COCO: common objects in context. In: European conference on computer vision. Springer, pp 740−755
79. Cai H, Wu Q, Hall P (2015) Beyond photo-domain object recognition: benchmarks for the cross-depiction problem. In: Proceedings of the IEEE international conference on computer vision workshops, pp 1−6
80. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579−2605
81. Crowley EJ, Zisserman A (2016) The art of detection. In: European conference on computer vision. Springer, pp 721−737
82. Gonthier N, Gousseau Y, Ladjal S, Bonfait O (2018) Weakly supervised object detection in artworks. In: Proceedings of the European Conference on Computer Vision (ECCV)
83. Ufer N, Lang S, Ommer B (2020) Object retrieval and localization in large art collections using deep multi-style feature fusion and iterative voting. In: European conference on computer vision. Springer, pp 159–176
84. Pease A, Colton S (2011) On impact and evaluation in computational creativity: a discussion of the Turing test and an alternative proposal. In: Proceedings of the AISB symposium on AI and Philosophy, vol 39
85. Tan WR, Chan CS, Aguirre HE, Tanaka K (2017) ArtGAN: artwork synthesis with conditional categorical GANs. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp 3760−764
86. Tan WR, Chan CS, Aguirre HE, Tanaka K (2018) Improved ArtGAN for conditional synthesis of natural image and artwork. IEEE Trans Image Process 28(1):394−409
87. Lin M, Deng Y, Tang F, Dong W, Xu C (2020) Multi-attribute guided painting generation. In: 2020 IEEE conference on

multimedia information processing and retrieval (MIPR). IEEE, pp 400–403
88. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4401−4410
89. Jin Y, Zhang J, Li M, Tian Y, Zhu H, Fang Z (2017) Towards the automatic anime characters creation with generative adversarial networks arXiv preprint arXiv:1903.026786
90. Liu L, Zhang H, Xu X, Zhang Z, Yan S (2019) Collocating clothes with generative adversarial networks cosupervised by categories and attributes: a multidiscriminator framework. IEEE Trans Neural Netw Learn Syst 31(9):3540−3554
91. Tomei M, Cornia M, Baraldi L, Cucchiara R (2019) Art2Real: unfolding the reality of artworks via semantically-aware image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5849−5859
92. Tomei M, Cornia M, Baraldi L, Cucchiara R (2019) Image-to-image translation to unfold the reality of artworks: an empirical analysis. In: International conference on image analysis and processing. Springer, pp 741−752
93. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2414−2423
94. Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M (2019) Neural style transfer: a review. IEEE Trans Vis Comput Graph 36:3365−3385
95. Elgammal A, Kang Y, Den Leeuw M (2018) Picasso, Matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018. AAAI Press, pp 42−50
96. Deng Y, Tang F, Dong W, Ma C, Huang F, Deussen O, Xu C (2020) Exploring the representativity of art paintings. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2020.3016887
97. Lu X, Sawant N, Newman MG, Adams RB, Wang JZ, Li J (2016) Identifying emotions aroused from paintings. In: European conference on computer vision. Springer, pp 48−63
98. Cetinic E, Lipic T, Grgic S (2018) How convolutional neural networks remember art. In: 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, pp 1−5
99. Cetinic E, Lipic T, Grgic S (2019) A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access 7:73694−73710
100. Isola P, Xiao J, Torralba A, Oliva A (2011) What makes an image memorable?. In: CVPR 2011. IEEE, pp 145−152
101. Li D, Yang Y, Song Y-Z, Hospedales T (2018) Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
102. Carlucci FM, D'Innocente A, Bucci S, Caputo B, Tommasi T (2019) Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2229−2238
103. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618−626
104. Jetley S, Lord NA, Lee N, Torr PH (2018) Learn to pay attention. arXiv preprint arXiv:1804.02391
105. Costa V, Dellunde P, Falomir Z (2021) The logical style painting classifier based on Horn clauses and explanations (l-SHE). Log J IGPL 29(1):96−119
106. Aggarwal G, Parikh D (2020) Neuro-symbolic generative art: a preliminary study. arXiv preprint arXiv:2007.02171
107. Amizadeh S, Palangi H, Polozov O, Huang Y, Koishida K (2020) Neuro-symbolic visual reasoning: disentangling visual from reasoning. arXiv preprint arXiv:2006.11524
108. Mercuriali G (2019) Digital art history and the computational imagination. Int J Digit Art Hist Issue 3 2018 Digit Space Architect 3:141
109. Trejo K, Angulo C, Satoh S, Bono M (2018) Towards robots reasoning about group behavior of museum visitors: leader detection and group tracking. J Ambient Intell Smart Environ 10(1):3−19
110. Castellano G, Carolis BD, Macchiarulo N, Vessio G (2020) Pepper4Museum: towards a human-like museum guide. In: Antoniou A, et al (eds) Proceedings of the AVI2CH workshop on advanced visual interfaces and interactions in cultural heritage, co-located with 2020 International Conference on Advanced Visual Interfaces (AVI 2020), vol 2687, CEUR-WS, 28 September−2 October 2020