IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 21, 2019, accepted May 31, 2019, date of publication June 5, 2019, date of current version June 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921101

A Deep Learning Perspective on Beauty, Sentiment, and Remembrance of Art

EVA CETINIC[1], TOMISLAV LIPIC[1], AND SONJA GRGIC[2], (Member, IEEE)
[1]Rudjer Boskovic Institute, 10 000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, 10 000 Zagreb, Croatia

Corresponding author: Eva Cetinic (ecetinic@irb.hr)

This work was supported by the European Regional Development Fund under Grant KK.01.1.1.01.0009 (DATACROSS).

ABSTRACT With the emergence of large digitized fine art collections and the successful performance of deep learning techniques, new research prospects unfold in the intersection of artificial intelligence and art. In order to explore the applicability of deep learning techniques in understanding art images beyond object recognition and classification, we employ convolutional neural networks (CNN) to predict scores related to three subjective aspects of human perception: aesthetic evaluation of the image, sentiment evoked by the image, and memorability of the image. For each concept, we evaluate several different CNN models trained on various natural image datasets and select the best performing model based on the qualitative results and the comparison with existing subjective ratings of artworks. Furthermore, we employ different decision tree-based machine learning models to analyze the relative importance of various image features related to the content, composition, and color in determining image aesthetics, visual sentiment, and memorability scores. Our findings suggest that content and image lighting have significant influence on aesthetics, in which color vividness and harmony strongly influence sentiment prediction, while object emphasis has a high impact on memorability. In addition, we explore the predicted aesthetic, sentiment, and memorability scores in the context of art history by analyzing their distribution in regard to different artistic styles, genres, artists, and centuries. The presented approach enables new ways of exploring fine art collections based on highly subjective aspects of art, as well as represents one step forward toward bridging the gap between traditional formal analysis and the computational analysis of fine art.

INDEX TERMS Convolutional neural networks, image aesthetics, image memorability, fine art, visual sentiment.

I. INTRODUCTION

Deep learning techniques have been successfully employed for resolving a wide variety of tasks in many different areas. With the rise of digitized and online available fine art collections, new perspectives emerge for employing deep learning techniques within the art domain. In particular, convolutional neural networks currently outperform all other computational methods for the task of classifying paintings by artist, style or genre. Apart from solving the challenge of automatic classification of artworks, deep neural networks have the potential to enable new ways of exploring digitized art collections, as well as to discover new patterns and meaningful relations among specific artworks or artistic oeuvres. Fine art collections are a data source of historically relevant, as well as perceptually and emotionally intriguing visual information. Because of its manifold nature, the domain of fine art images represents a fruitful data source for formulating semantically relevant image analysis tasks, as well as for challenging neural networks in learning representations of a higher abstraction level.

In order to explore the applicability of convolutional neural networks in understanding images beyond object detection and classification, we aim to address image properties related to the subjective and affective aspects of human perception. We focus on three different levels of perceiving images: the aesthetic evaluation of the image; the sentiment evoked by the image and the memorability of the image. These three different aspects of image perception have been studied by psychologists for a long time [1]–[4] and have recently become an emerging subject of interest within the computer vision and machine learning community. Due to the higher

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

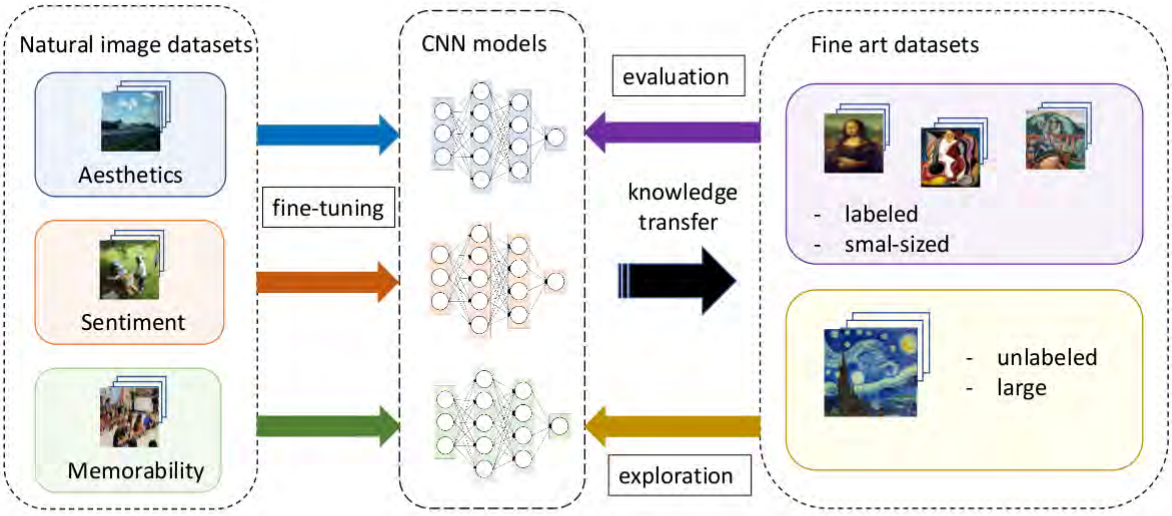

FIGURE 1. Conceptual overview of utilizing knowledge transfer of aesthetic, sentiment and memorability prediction from natural images to fine art.

availability of natural image datasets, most of the studies concerned with computationally addressing perceptual image features were done on photographs, while art images have not yet been systematically explored. The appearance of big and comprehensively annotated fine art datasets facilitates the analysis of those specific visual properties on a large scale. However, collecting ground-truth labels for attributes related to subjective perception of images is laborious and expensive because it requires complex experimental surveys. On the other hand, the concept of transfer learning and the transferability of pre-trained CNN models across different domains enable new ways of feature assessment.

In this work we employ several CNN models trained to predict aesthetic, sentiment and memorability scores of natural images in order to explore those features in art images. Fig. 1 illustrates the methodology used in this study. Various CNN models trained on different natural image datasets are evaluated on available small-sized annotated fine art dataset. Based on the correlation between predicted scores and human judgments, we select the best performing CNN model for each task for further analysis. We particularly focus on how the predicted aesthetics, sentiment and memorability scores relate to different artistic styles and genres, as well as how they correlate with each other and various visual attributes.

The main contributions of our work are:

- Deep learning based quantitative approaches to highly subjective aspects of perceiving images (aesthetics, memorability, sentiment) are employed for the first time in the domain of fine art images.
- The presented approach of utilizing knowledge transfer by employing CNN models trained for predicting aesthetic, sentiment and memorability scores in natural images to fine art images generates meaningful qualitative and quantitative results.
- Several different CNN models trained on various natural image datasets are evaluated by comparing the predicted results with subjective ratings available for several small-scale fine art datasets.
- Global exploratory analysis of a large-scale fine-art collection is performed in order to study the relation of predicted aesthetics, sentiment and memorability scores with each other and with other high-level image attributes within the domain of fine art.
- Analysis of the distribution of predicted aesthetic, sentiment and memorability scores is performed in regard to different artistic styles, genres, artists and centuries.

II. RELATED WORK

We present the research related to our work by giving a summary of four research directions that interweave in our work. First we focus on computational analysis of art by presenting a summary of relevant approaches that apply different computer vision and machine learning techniques to the domain of fine art images. Additionally, we provide short overviews of research related to computational aesthetics, visual sentiment analysis and image memorability estimation.

A. COMPUTATIONAL ANALYSIS OF ART

The emergence of large digitized and online available fine art collections facilitated the opening of new research questions in the interdisciplinary field of computer vision, machine learning and art history. Analyzing artworks includes understanding different aspects such as form, expression, content and meaning. Those aspects arise from formal elements of paintings such as line, shape, color, texture and composition [5]. Various computational image features can be used in order to analyze and describe the formal elements of art images, primarily color or texture-based features. However, narrowing the semantic gap between low-level image features

and artistic concepts remains a great challenge in computational analysis of art.

The majority of studies concerned with computational analysis of art images is focused on the challenge of automatically classifying artworks based on categories such as artist, style or genre. Most of the earlier studies addressing the topic of automatic artist classification [6], [7], as well as style [8] and genre classification [9] are based on extracting a set of various low-level image features and using them to train different classifiers. Recently significant progress of classification performance has been achieved with the adoption of convolutional neural networks. Karayev et al. introduced the approach to use layers activations of a CNN trained on ImageNet [10], a large hand-labeled object dataset, as features for artistic style recognition [11]. They showed that features derived from an object recognition pre-trained CNN outperform most of the hand-crafted features on the task of style classification. The advantage of CNN-based features, particularly in combination with other hand-crafted features, was confirmed for artist [12], style [13] and genre classification [14]. Besides using pre-trained CNNs just as feature extractors, Girshick et al. showed that further improvement of performance for a variety of visual recognition tasks can be achieved by fine-tuning a pre-trained network on the new target dataset [15]. The predominance of this approach has been confirmed for various classification task on artistic datasets as well [16]–[19].

Apart from classification, the use of CNN-based features showed promising results in other topics of interests such as retrieving visual links in paintings collections [20], recognizing objects in paintings [21], [22] or distinguishing illustrations from photographs [23]. Recently there has been an emerging tendency towards enhancing the interpretability of learned representations, as well as understanding their relation to domain-specific properties and their position in a wider semantic context. Elgammal *et al.* [24] performed a correlation analysis of learned features extracted from several different style-trained CNN models in order to understand how learned representations are related to art history methodologies for identifying styles. Furthermore, Brachmann *et al.* [25] used CNN features to understand the specific visual properties of artworks in comparison to natural images.

B. COMPUTATIONAL AESTHETICS

Computational aesthetics is a growing field of interest within the computer vision community and is mainly preoccupied with developing computational methods that can predict aesthetic judgments in a similar manner as humans. Although some studies have addressed the topic of computational aesthetics in art by analyzing the correlation between statistical image properties and Western paintings [26] or computationally evaluating the aesthetics of Chinese calligraphy [27], [28], the majority of studies in this field focuses on predicting aesthetic rating of photographs. As in many other computer vision related tasks, this was first done by extracting various low level hand-crafted image features to train different types of classifiers [29], [30], usually using datasets annotated with subjective aesthetic ranking scores obtained through different experimental surveys [31]. In more recent works, the adoption of learned deep features showed significant improvement [32], particularly when fine-tuning pre-trained CNN models for the task of predicting aesthetic scores [33]–[35]. For a more detailed overview of studies in computational aesthetics, we refer the reader to [36], [37].

C. VISUAL SENTIMENT ANALYSIS

With the increasing of online visual data, understanding sentiment in visual media is gaining more and more research attention. Most commonly, research activities revolve around two different directions: 1) recognizing the sentiment expressed through facial expressions and bodily gestures depicted in the image; 2) detecting the sentiment that particular image content and visual properties evoke in human observers. In our work we consider the second direction of visual sentiment analysis, particularly in regard to sentiment polarity and the determination of whether an image expresses positive or negative sentiment. For this purpose various methods have been developed over the years. In most previous works, a common approach was to correlate low-level image features with high-level visual attributes [38] and train a classifier using human annotations as ground truth [39]. Recent methods based on the employment of CNNs demonstrate superiority in predicting visual sentiment [40], particularly when fine-tuning CNN models using human-annotated datasets [41], [42].

D. IMAGE MEMORABILITY ESTIMATION

Image memorability is a concept that refers to how easy it is for a person to remember a certain image. Image memorability has been studied by psychologists for a long time [2], [4] and it has been shown that people tend to remember the same kind of images. This indicates that the phenomenon of memorability exceeds the mere subjective experience and that certain visual properties are universally more memorable than others. Recently, image memorability became a subject of interest within the computer vision community when Isola *et al.* [43] developed a framework for predicting image memorability based on global image descriptors. They built a human-annotated dataset by collecting responses through a visual memory game and trained a support vector regression (SVR) model to map different hand-crafted image features into memorability scores. Additionally, they analyzed the correlation between specific image features and memorability. Their results indicated that simple image features do not correlate strongly with memorability and that content has a significant impact on memorability, with photos of people being more memorable than photos of landscapes. Following their work, other approaches were proposed to improve memorability prediction by investigating different image features [44], [45]. A comprehensive overview of studies related to image memorability is given in [46]. The adoption of

TABLE 1. List of all datasets used in this work. For each dataset we indicate the phase in which it was used, the corresponding task, the number of images, the number of collected ratings per image, and the type of images in the dataset: Fine art images (F) or natural images (N).

Phase	Task	Dataset	# images	# ratings per image	Type
Exploration	Aesthetic Sentiment Memorability	WikiArt	105 121	-	F
Evaluation	Aesthetic	JenAesthetic [49]	1568	∼20	F
Evaluation	Sentiment	MART [50]	500	20	F
Evaluation	Sentiment	WikiEmotions [51]	3379	∼10	F
Training	Aesthetic	AADB [34]	10 K	5	N
Training	Aesthetic	AVA [31]	250 K	∼210	N
Training	Aesthetic	FLICKR-AES [52]	40 K	5	N
Training	Sentiment	Twitter DeepSent [41]	1269	- 5 agree	N
Training	Sentiment	Flickr Sentiment [53]	10 K	- 3 agree	N
Training	Memorability	LaMem [47]	58 741	∼80	N
Training	Memorability	SUN Memorability [54]	2222	∼78	N

CNNs for the task of image memorability was introduced by Khosla *et al.* [47] and advanced in more recent studies [48].

III. METHODOLOGY

CNNs have become very popular for solving a variety of different image recognition and classification tasks. One of the main reasons for the breakthrough of deep CNNs was the availability of large hand-labeled object categorization datasets such as ImageNet [10]. The purpose of deep CNN models trained on ImageNet went beyond the initial object classification tasks when it became evident that fine-tuning models pre-trained on ImageNet data, using smaller datasets and for different tasks, yielded state-of-the-art results for various image classification tasks. Fine-tuning CNNs has become a common practice for solving many computer vision tasks and has resulted in a growing collection of pre-trained CNN models. Those models represent a valuable source for transferring knowledge across different domains and applications. Several recent works showed that reusing models trained for natural images can produce remarkably good results for object recognition in paintings [21], [22], even without fine-tuning or domain adaptation. To tackle the applicability of pre-trained CNNs beyond the cross-depiction problem, we aim to explore if models pre-trained on natural images can extract perceptually-related images features when applied on fine art images.

Our proposed methodology can be divided into six main steps. In the first step, we employ 3 different CNN models trained on natural images for each task (9 models in total). In order to have 3 different models for each task, we collect pre-trained models made available by others as well as fine-tune new models using different natural image datasets. In the second step, we compare the predicted scores obtained by three different CNN models on the large unlabeled fine art dataset in order to investigate the consistency of predictions for each task. In order to evaluate the results and choose the best performing model for each task, in the third step we employ the models on available small-sized annotated fine art dataset and compare the CNN predicted scores with human evaluation scores. After identifying the best performing model for each task based on the correlation between the predicted scores and human rating scores, in the fourth step we evaluate the qualitative results of the predicted aesthetic, sentiment and memorability scores on the large unlabeled fine art dataset by visually inspecting the images with the highest and lowest prediction values. In the fifth step we analyze the correlation of the predicted scores with other image features. In the final step, we provide an analysis of the predicted aesthetic, sentiment and memorability scores in relation to styles, genres, artists and time frames.

IV. EXPERIMENTAL SETUP

In this section we provide details regarding the image dataset and the pre-trained CNN models used in the experiments.

A. DATASETS

In this section we give a description of the all fine art and natural image datasets used in this work. The various datasets were used for three different phases: (1) to explore the correlation between different concepts in the domain of fine art images; (2) to evaluate the machine-based predictions with human judgments of fine art images and (3) to train deep neural networks for the tasks of aesthetic, sentiment or memorability prediction. The datasets are listed in Table 1.

To study the correlation between different concepts in the domain of fine art images, we collect images from WikiArt.org. To the best of our knowledge, the WikiArt dataset is currently the largest online available fine art dataset, as well as the most commonly used dataset for automated classification tasks. It includes artworks from a wide time period, with a large corpus of 19th and 20th century paintings, as well as contemporary art. The WikiArt collection includes images annotated with a large set of labels such as artist, genre, style, technique, etc. At the time of our data

TABLE 2. List of all CNN models used in this work. For each model we indicate the task for which it was trained, the model ID, the type of architecture, the dataset that was used for training the model and the source of the model.

Task	Model ID	Architecture	Dataset	Source
Aesthetics	AestNet_1	AlexNet with attribute branches	AADB	[34]
Aesthetics	AestNet_2	GoogLeNet	AVA	[35]
Aesthetics	AestNet_3	ResNet50 + soft attention + LSTM	FLICKR-AES	ours
Sentiment	SentiNet_1	AlexNet	Twitter	[42]
Sentiment	SentiNet_2	AlexNet	Flickr Sentiment	ours
Sentiment	SentiNet_3	ResNet50 + soft attention + LSTM	Flickr Sentiment	ours
Memorability	MemNet_1	AlexNet	LaMem	[47]
Memorability	MemNet_2	ResNet50 + soft attention + LSTM	SUN Memorability	[48]
Memorability	MemNet_3	ResNet50 + soft attention + LSTM	LaMem	[48]

collection process, this dataset contained more than 130K images of various artworks. However, we decided to use only paintings and therefore excluded artworks that were classified as photography, poster, architecture, graffiti, installation, etc. In addition, in order to include only color images, we removed all grayscale prints and created a final subset of 105 121 images.

In order to explore the transferability of learned aesthetic, sentiment and memorability features from the domain of natural images to the domain of fine art images, we employ several different CNN models that were trained on different domain-specific natural image datasets. Based on the availability and quality of existing models and datasets, we identify different datasets for each task. For the purpose of aesthetic quality prediction, we employ models that were trained on three different datasets: AADB [34] , AVA [31] and FLICKR-AES [52]. For visual sentiment classification we employ models that were trained on two different datasets: Twitter DeepSent [41] and Flickr Sentiment [53]. For memorability prediction, we use models trained on the LaMem [47] and SUN Memorability [54] datasets.

Collecting ground-truth labels for subjective attributes such as aesthetic, sentiment and memorability of images is complex and expensive. Although there are several large-scale annotated natural image datasets for all three tasks, only a few small-sized fine art datasets are available for the tasks of aesthetic and sentiment prediction and none for memorability. We use those available fine art datasets in order to assess the performance of CNN-based predictions in relation to human judgments of aesthetics and sentiment in fine art paintings. Based on the Spearman's rank correlation coefficient between the predicted and ground-truth aesthetic and sentiment scores, we decide which model should be employed for predicting aesthetic and sentiment scores for the exploratory analysis on the WikiArt dataset.

For evaluating aesthetic scores, we use the JenaAesthetic dataset [49], [55]–[57]. The dataset contains images of 1568 different oil paintings by 410 artist from 11 different art periods. The labels were collected by asking participants to rate different properties of the images such as "aesthetic quality", "beauty", "color", "content" and "composition". Each painting was rated by 19 to 21 observers and the median value between the individual scores is considered as the ground-truth value for each property.

For evaluating the predicted sentiment scores we use the MART dataset which consists of 500 abstract paintings [50] labeled as evoking positive or negative sentiment. Each artwork received 20 ratings from 20 different subjects and the average score is considered the ground-truth value. Besides evaluating the predicted sentiment scores on the MART dataset which includes positive-negative emotional judgments, we use the newly introduced WikiArt Emotions dataset [51] to analyze how the predicted sentiment scores relate to different emotion categories. The WikiArt Emotions dataset includes emotion annotations for 4105 artworks, belonging to 22 different style categories, collected from WikiArt.org. For the purpose of our work we use a subset of 3379 paintings and analyze the relationship between the predicted sentiment scores and 8 different emotion categories.

B. CNN MODELS

In this section we give a brief description of the all CNN models employed for the purpose of extracting aesthetic, sentiment and memorability prediction scores, as well as high-level image attributes.

1) AESTHETICS, SENTIMENT AND MEMORABILITY SCORES

For obtaining aesthetic, sentiment and memorability prediction scores we select several different CNN models for each task in order to analyze the consistency and correlation between outputs of different models. In Table 2 we provide a list of all models, including information about the type of architecture and the dataset used for training. In order to have at least three models for each task, we employ both models made available by others, as well as train our own models. The Source column in Table 2 indicates whether the model is introduced by others or is fine-tuned for the purpose of this work.

Introduced by Kong *et al.* [34], AestNet_1 is based on the AlexNet architecture and is fine-tuned using the "Aesthetics

with Attributes Database'' (AADB), which contains aesthetic scores and high-level visual attributes assigned to each image by multiple human raters. The original AlexNet softmax classification layer is replaced with an Euclidean Loss regression layer and attribute prediction branches are added on top of the second fully connected layer. The predicted aesthetic scores show a high level of consistency with human ratings on the AADB dataset. Originally named ILGNet, the AestNet_2 model was introduced by Xin Jin *et al.* [35]. This is a GoogLeNet model pre-trained for image classification and fine-tuned for image aesthetic quality classification on the AVA [31] dataset, achieving a classification accuracy of 81.68%. The third model, AestNet_3 is based on the AMNet architecture introduced in [48]. The AMNet architecture consists of an ImageNet pre-trained ResNet50 model, followed by a modified visual attention mechanism with a Long Short Term Memory (LSTM) recurrent network [58] and a network for regression. The AMNet network was originally introduced for the purpose of image memorability estimation but was designed in a generic manner so it could be employed for other regression tasks. For the purpose of estimating the aesthetic quality of images, we fine-tune the AMNet network on the FLICK-AES dataset [52] and achieve a Spearman's rank correlation coefficient of 0.72 between the predicted and ground-truth ranking.

Similarly, for the purpose of sentiment estimation, we fine-tune the AMNet network on the Flicker Sentiment dataset introduced in [53] and achieve a Spearman's rank correlation coefficient of 0.53. We refer to this model as the SentiNet_3 model. In addition to this model, we introduce the SentiNet_2 model as a result of fine-tuning an ImageNet pre-trained CaffeNet model on a subset of the Flickr Sentiment dataset. This model is fine-tuned to classify images as evoking either positive or negative sentiment. As the task distinguishes two classes of positive and negative sentiment, a two-neuron layer replaces the original fc8 layer in CaffeNet. The SentiNet_2 model achieves a classification accuracy of 0.84% on the Flicker-Sentiment dataset. Additionally, we employ the SentiNet_1 model which is the result of a similar approach introduced by Campos *et al.* [42], where a CaffeNet model was fine-tuned on the Twitter DeepSent dataset [41].

For the task of predicting the memorability scores of images, we also employ three different models. MemNet_1 refers to the model introduced by Khosla *et al.* [47]. This model is a result of fine-tuning a pre-trained CaffeNet model on the LaMem dataset. LaMem is a large memorability dataset, consisting of 60 000 images annotated with human memory scores conducted through a memory game experiment. The model achieved a Spearman's rank correlation of 0.64, with 0.68 being the human consistency rank correlation. An improvement on this results was achieved in [48] with the original AMNet model on the Lamem dataset (MemNet_2) and the SUN dataset (MemNet_3).

2) HIGH-LEVEL IMAGE ATTRIBUTES

To better understand how the predicted aesthetic, sentiment and memorability scores relate to different image properties, we analyze their correlation with different high-level image attributes. Inspired by traditional fine art and photographic principles, high-level image attributes represent interpretable characteristics of content, composition and color. Various hand-crafted features have been proposed [30], [38] for the purpose of quantifying different high-level attributes. However, designing features that capture high-level attributes is a difficult task and using learned features instead of engineered features has shown to be better for a variety of computer vision applications. As the ability of CNN models to automatically identify meaningful patterns has proven useful for learning complex image attributes [33], [34], we employ the aesthetic model proposed by Kong *et al.* [34] to extract high-level image attributes. The model is trained on the AADB dataset, where each image is annotated with an aesthetic quality rating and attribute assignments provided by five different individual raters. A confidence score is assigned to each attribute based on the aggregation of assignments by multiple raters. The CNN model implements an attribute prediction task that can be viewed as a source of side information which serves to regularize the weights during training, but is not part of the test aesthetic score prediction. However, for a given input image, each attribute layer can be employed to output a prediction score which quantifies the intensity of a specific attribute. For that reason, attribute layers can be considered as feature extractors that are independent of the predicted aesthetic score. We use outputs of the attribute layers in order to obtain scores with values in the range from 0 to 1. The values indicate the extent to which an attribute is present in the image. In order to support the choice of this model for extracting high-level attributes with qualitative results, we provide figures showing artworks with the top 100 highest and top 100 lowest scores for each attribute in the Supplemental files (Fig. S1 - S8). We selected the following eight attributes: content (if the image has positively rated content), object emphasis (if foreground objects are emphasized in the image), lighting (if the image has good lighting), rule of thirds (if the image composition follows the rule of thirds), repetition (if the image has repetitive patterns), symmetry (if the image is symmetric), color harmony (if the overall colors are combined in a harmonious way), color vividness (if the colors are bright and intense).

V. RESULTS

The experimental results are presented from several viewpoints. First we analyze the attention maps of different task-specific CNN models, as well as compare and evaluate predicted aesthetic, sentiment and memorability scores. Furthermore, we explore the relation of different image attributes with the predicted aesthetic, sentiment and memorability scores. Finally, we position the results in the context

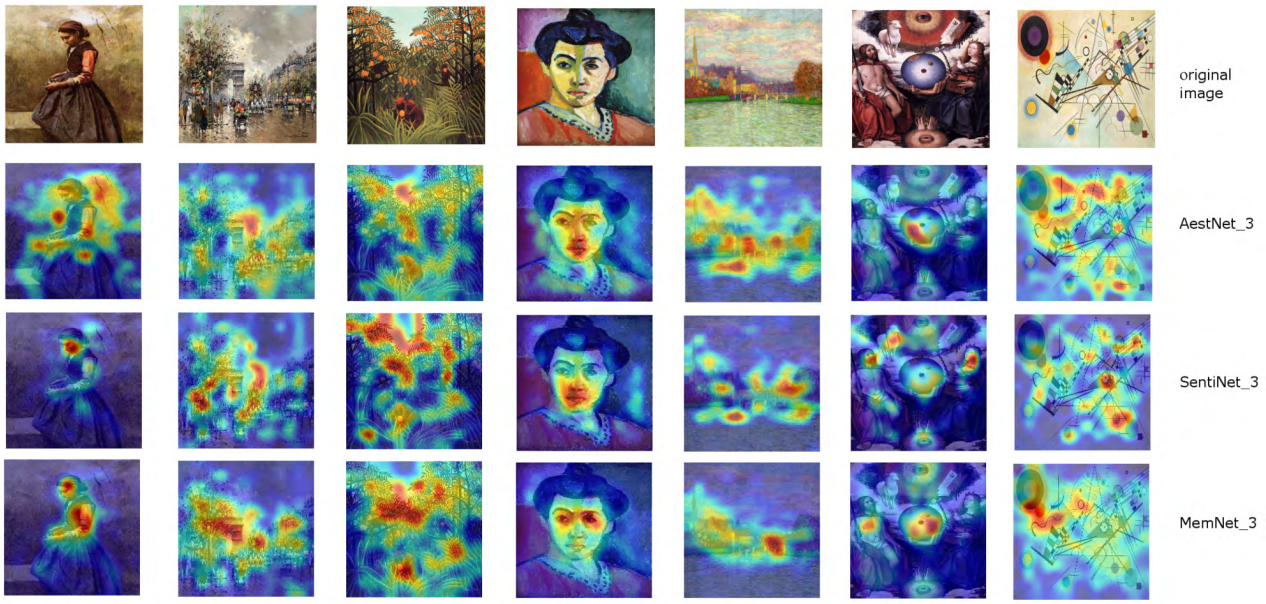

FIGURE 2. Examples of attention maps employed for predicting the aesthetic score (AestNet_3), sentiment score (SentiNet_3) and memorability score (MemNet_3).

of art history and investigate their relation to style, genre, artists, as well as how they change over time.

A. ANALYSIS OF ATTENTION MAPS

To understand how models trained for different tasks process the same image, we compare attention maps obtained from the three models trained with the soft attention mechanism (AestNet_3, SentiNet_3 and MemNet_3). The architecture of these models is introduced and described in detail in [45]. To obtain attention maps, each image is represented with an $14\times14\times1024$ tensor which is the output of the 43rd layer of a RestNet50 network trained for image classification on the ImageNet dataset. In this image representation there are 14×14 image locations associated with corresponding 1024 dimensional feature vectors. The soft attention mechanism produces a probability weight for every image location. The soft attention probabilities are conditioned on the entire image feature vector and the previous LSTM hidden state and represented as a vector of weights produced by a softmax function. The output of the softmax function is scaled to range [0, 255] and resized from 14×14 to 244×244 in order to obtain images of attention maps with the same resolution as the original input image. The prediction score of a particular task (aesthetic, sentiment or memorability) is estimated with LSTM over a three steps long sequence. In the subsequent LSTM steps, the attention moves to the regions responsible for estimating the scores of a particular task. Fig. 2 shows examples of attention maps for different fine art images produced in the final LSTM step.

The attention maps for aesthetic prediction tend to cover larger regions of images, while sentiment and memorability usually localize into few smaller peaks. Face and body regions are commonly triggered for sentiment and memorability. In addition to the attention maps shown in Fig. 2, in the Supplemental files (Fig. S9) a comparison of averaged attention maps over all three LSTM steps for 300 images with largest scores and 300 images with smallest scores in each of the three most represented genres (abstract, landscape, portrait) in WikiArt dataset.

B. QUANTITATIVE EVALUATION OF AESTHETIC, SENTIMENT AND MEMORABILITY SCORES

Using the images in the WikiArt dataset as inputs to the pre-trained models listed in Table 2, we collect three aesthetic, three sentiment and three memorability scores for each image. The output of each model is a value between 0 and 1. For aesthetic prediction, a higher output means that the model predicts a higher aesthetic evaluation of the image. In the case of sentiment prediction, a higher output value indicates that the image evokes positive sentiment, while lower values indicate negative sentiment. Similarly, in the case memorability estimation, higher output values indicate that the image is more memorable.

To analyze the consistency and relation between scores predicted by different models trained on the same task, we use Spearman's rank correlation coefficient, which indicates the strength and direction of the monotonic relationship between two variables. Fig. 3 shows the correlation between the scores obtained on the WikiArt using different models trained for the same task.

The correlation between outputs of different models is the strongest for the memorability task, while the different aesthetics scores have the weakest correlation. This demonstrates a stronger consistency of the results obtained from

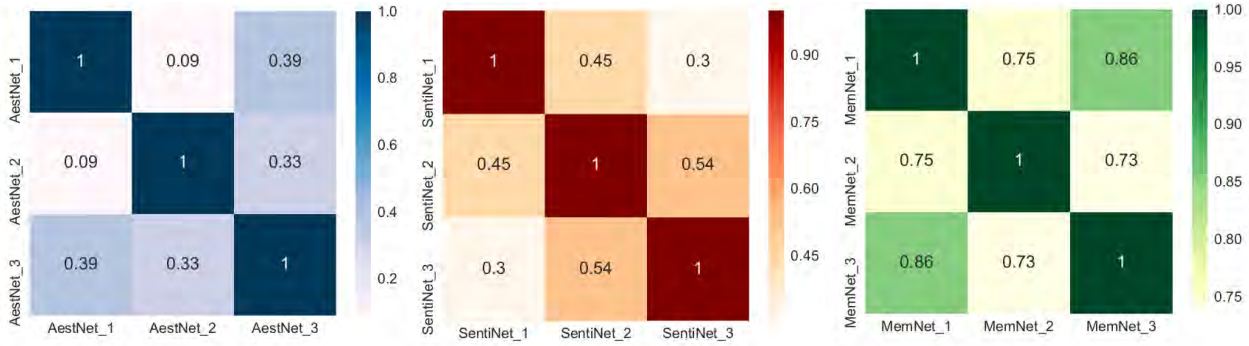

FIGURE 3. Heatmaps of spearman's correlation coefficients between the different aesthetic scores (left), sentiment scores (middle) and memorability scores (right) on the WikiArt dataset (p-value < 0.001).

TABLE 3. Values of the Spearman's rank correlation coefficient between predicted aesthetic scores of three different CNN models and the average of subjective scores for different properties on the JenAesthetic dataset (* p-value < 0.01, ** p-value > 0.1).

	Aesthetic quality	Color liking	Composition liking	Content liking
AestNet_1	0.040**	0.203*	0.029**	-0.003**
AestNet_2	0.166*	0.084*	0.131*	-0.076*
AestNet_3	0.267*	0.372*	0.287*	0.267*

TABLE 4. Values of the spearman's rank correlation coefficient between predicted sentiment scores of three different CNN models and the the proportion of different emotion annotations for each image in the WikiEmotion dataset ((p-values < 0.01), except for *(p < 0.1)).

	Sadness	Disgust	Fear	Pessimism	Happiness	Love	Optimism	Trust
SentiNet_1	-0.201	-0.024	-0.253	-0.195	0.056	0.028*	0.106	-0.054
SentiNet_2	-0.205	-0.086	-0.252	-0.202	0.230	0.061	0.223	0.004
SentiNet_3	-0.142	-0.217	-0.205	-0.178	0.396	0.246	0.268	0.289

different models trained for image memorability prediction. We presume that there is a weaker correlation between outputs of different models trained for predicting aesthetic evaluation because aesthetic evaluation is generally more subjective than sentiment or memorability and thus more difficult to automatically predict. Also, these inconsistencies may arise from different experimental setups used for the preparation of different aesthetic datasets (number of raters, differences in the choice of images, differences in the questionnaires, etc.).

1) AESTHETICS

In order to select the model which would be most suitable for performing a large-scale exploratory analysis on the WikiArt dataset, we use the small-sized fine art datasets annotated with human judgments for evaluation. The Spearman's rank correlation coefficients between the predicted aesthetic scores and the ground-truth scores of images in the JenAesthetic dataset are presented in Table 3 .

The results indicate that the aesthetic scores predicted using the AestNet_3 model have the highest correlation coefficient with the ground-truth aesthetic quality ratings on the JenAesthetic dataset, as well as with ratings of properties related to the evaluation of color, composition and content. The AestNet_2 scores have a weak positive correlation with the aesthetic quality and composition ratings, but no significant correlation with other properties, while the AestNet_1 scores are positively correlated only with the subjective ratings of color. Based on these results, we select the AesNet_3 model for studying the relation between predicted aesthetic scores and other image attributes, as well as contextual categories in the WikiArt dataset.

2) SENTIMENT

For the purpose of assessing the sentiment scores of different models, we study the consistency of their performance on two different datasets. Firstly, we analyze the values of the Spearman's rank correlation coefficient between predicted sentiment scores and the valance ratings of images in the MART dataset. Our analysis shows that predicted scores of all three models have a significant positive correlation with the ground-truth scores. Specifically, the SentiNet_1 scores have a correlation coefficient $\rho = 0.531$, SentiNet_2 scores $\rho = 0.546$ and SentiNet_3 scores $\rho = 0.483$ (p-value < 0.01 in all cases) with the ground-truth values of images in the MART dataset. Additionally, we employ all three models on the WikiEmotion dataset in order to explore how the positive sentiment scores correlate with different categorical representations of emotions. In Table 4 we report the Spearman's correlation coefficients between the predicted

FIGURE 4. Aesthetic - top 100 artworks with the highest (left) and the lowest (right) values of predicted aesthetic scores.

sentiment scores and the proportion of different emotion annotations for each image in the WikiEmotion dataset. Although SentiNet_2 scores have the highest correlation with the ground-truth annotations on the exclusively abstract MART dataset, we select the SentiNet_3 model for further analysis because SentiNet_3 scores show a stronger consistency with the proportion of positive emotion annotations in the more diverse WikiEmotion dataset.

3) MEMORABILITY

Although there is no fine art dataset with ground-truth memorability scores for evaluating the performance of the different models trained for memorability, the high correlation between scores indicates a robustness of the memorability results. However, MemNet_3 achieved state-of-the-art memorability prediction performance on the LaMem dataset, attaining rank correlation of 0.677, with 0.68 being the human consistency rank correlation. Because of its near-human level of consistency in predicting memorability, this model is employed for obtaining scores for further analysis.

C. QUALITATIVE EVALUATION OF AESTHETIC, SENTIMENT AND MEMORABILITY SCORES

In this section we analyze the qualitative results by assembling images with the highest and lowest prediction values in order evaluate how the different CNN models predict scores based on the image content and visual features.

1) AESTHETICS

Regarding the prediction of the aesthetic evaluation, Fig. 4 shows fine art images with 100 highest (left) and 100 lowest (right) aesthetic scores predicted using AestNet_3. A short glimpse at the two embeddings already indicates that the major difference between the high and low valued images lies in color and lighting. Highly rated images tend to include bright and intense colors, while low rated images are dim and pale. This is an intuitively appropriate, although simplified notion of aesthetic evaluation, particularly regarding fine art images. However, the highly subjective nature of aesthetic experience and the consensus-based approach adopted in the model training process, sets limitations on the refinement of aesthetic criteria suitable for the context of fine art.

2) SENTIMENT

In order to understand properties which contribute to the evaluation of the positive or negative image sentiment, Fig. 5 shows fine art images with the 100 highest sentiment prediction scores (left) and 100 lowest sentiment prediction scores obtained with SentiNet_3. The obvious visual difference is the color choice, with positive images being bright and colorful and negative images darker. Regarding content, positively categorized images most commonly depict flowers, portraits of smiling people, couples and family portraits. Negatively categorized images often depict outdoor scenes, abstract images with sharp edges and strong contrast, as well as portraits of sad or fearful faces.

3) MEMORABILITY

When looking into the images with the highest and lowest memorability score obtained with MemNet_3, shown in Fig. 6, three dominant motifs occur within the most memorable images: abstract images with dot patterns, nude paintings and portraits. The least memorable images predominantly include outdoor scenes. This tendency is further confirmed in the analysis of the average memorability scores of artistic genres, presented in section V-E.2.

D. IMAGE FEATURES IN RELATION TO AESTHETIC, SENTIMENT AND MEMORABILITY SCORES

To better understand how different image properties correlate with the predicted aesthetic, sentiment and memorability scores, we analyze the correlation of different image features.

FIGURE 5. Sentiment - top 100 artworks with the highest (left) and the lowest (right) values of predicted sentiment scores.

FIGURE 6. Memorability - top 100 artworks with the highest (left) and the lowest (right) values of predicted memorability scores.

Particularly, we explore the relation with high-level features related to content, composition and color. For this purpose we use outputs of the attribute layers of the model proposed in [34], as described in Section IV-B.2. In order to observe how the predicted aesthetics, sentiment and memorability scores correlate with each other and other attribute scores, Fig. 7 shows Spearman's correlation matrix heatmap (on the right). The correlation is statistically significant because all p-values are close to zero. In addition, Fig. 7 shows pairwise linear relationships (upper triangle), estimated bivariate kernel densities (lower triangle) and corresponding univariate kernel densities of 8 image properties together with aesthetics, sentiment and memorability scores for all images in the WikiArt dataset (on the left).

The results indicate that aesthetics and sentiment scores are positively correlated ($\rho = 0.424$, p-value < 0.01), but both have a weak negative correlation with memorability. The predicted aesthetic score is moderately correlated with the content score, indicating that the content has a significant influence on aesthetics. Moreover, the aesthetic score positively correlates with lighting, as well as color harmony and vividness. On the contrary, memorability shows a weak negative correlation with color harmony, but is strongly positively correlated with object emphasis and negatively with repetition. This relation suggests that images depicting one salient object tend to be more memorable.

Although giving interesting insights about relations between high-level visual properties and aesthetics, sentiment and memorability scores, rank correlations only asses the strength of their pairwise monotonic relationship. In order to explore how different high-level image features jointly influence and predict aesthetics, sentiment and memorability scores, we employ different decision tree based regression machine learning models including single decision tree, in particular Classification and Regression Trees (CART) method [59], [60], as well as two popular tree-based

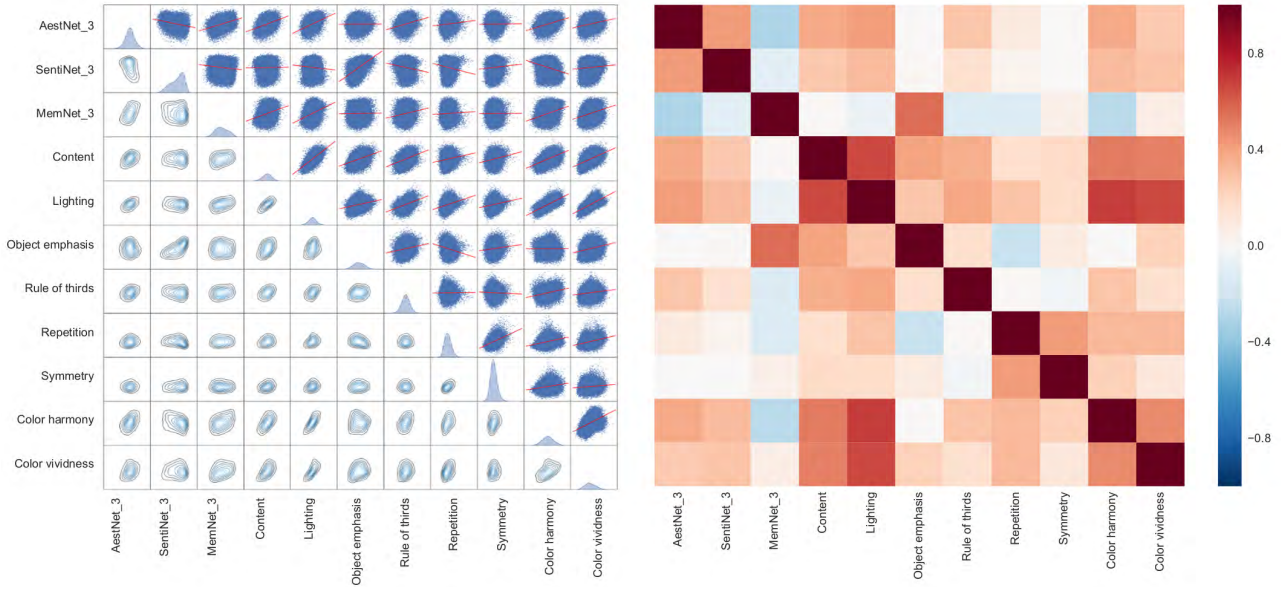

FIGURE 7. Pairwise relationships between 8 image properties and aesthetics, sentiment and memorability scores.

ensemble methods, random forests [61] and gradient boosting trees [62]. Specifically, we compare relative feature importances in determining image aesthetics, visual sentiment and memorability scores. The tree-based regression models automatically provide feature importances based on the contribution of different features to the score prediction. In our regression setting, this contribution is determined with respect to the reduction of mean-square error. It is important to note that when some of the correlated features are used in the model construction, the importance of other features is reduced. Three regression prediction tasks were formulated for image aesthetics, visual sentiment and memorability scores based on 8 high-level image features. The quality of regression models is evaluated by mean absolute error (MAE) between true and predicted score values. For each considered model in the regression tasks, different hyperparameter settings were explored using randomized search or Bayesian optimization (for gradient boosting trees). The model that yields the best averaged ten-fold cross validation mean absolute error was selected to train the final model on all available data. In addition, for a single CART-based decision tree, we also utilize greedy pruning strategy to obtain simple and more interpretable regression trees.[1]

The relative feature importances obtained from all models are shown in Fig. 8. Feature importances obtained from the AestNet_3, SentiNet_3 and MemNet_3 models indicate that the obtained results are robust with respect to the two most influential features for all regression tasks and models used. Specifically, features expressing the amount of lighting and positively rated content are dominant in determining aesthetics scores, while color vividness and harmony are crucial in influencing visual sentiment scores. Object emphasis clearly stands out as the most important feature for predicting memorability scores.

Because of the relative importance of color harmony and vividness in predicting aesthetic and sentiment scores, as well as the aim to analyze in more detail how specific colors influence the predicted perceptual features, we explore the correlation of aesthetic, sentiment and memorability scores with the amount of different hue values. We calculate a 12-bin normalized histogram of hue values in each image, where the position of bin edges corresponds to the 30 degrees interval of 12 major colors in the HSV color wheel. Fig. 9 shows the bar chart of Spearman's rank correlation coefficients for each of the 12 color subspaces. In order to show the consistency between hue names and image colors, the figure includes a small visualization of images retrieved as having the highest value for a particular hue category.

The correlation between the amount of a particular hue value and aesthetics, sentiment and memorability scores is not very strong. However, both aesthetic and positive sentiment scores have a weak negative correlation with red, while memorability is positively correlated only with red. The values of correlation coefficients between the amount of specific hues and predicted scores are similar for aesthetics and sentiment, but almost completely opposite for memorability. The hue correlations with aesthetic and sentiment scores are consistent with the general presumption that cool colors (green, cyan, blue) are preferred to warm colors (red, orange, yellow) [1], [3].

E. AESTHETIC, SENTIMENT AND MEMORABILITY IN THE CONTEXT OF ART HISTORY

This section provides an analysis of the predicted aesthetic, sentiment and memorability scores in relation to concepts related to art history. The section is divided based on

[1]Simplified tree-based classifier and regressor for interpretable machine learning, https://github.com/tmadl/sklearn-interpretable-tree

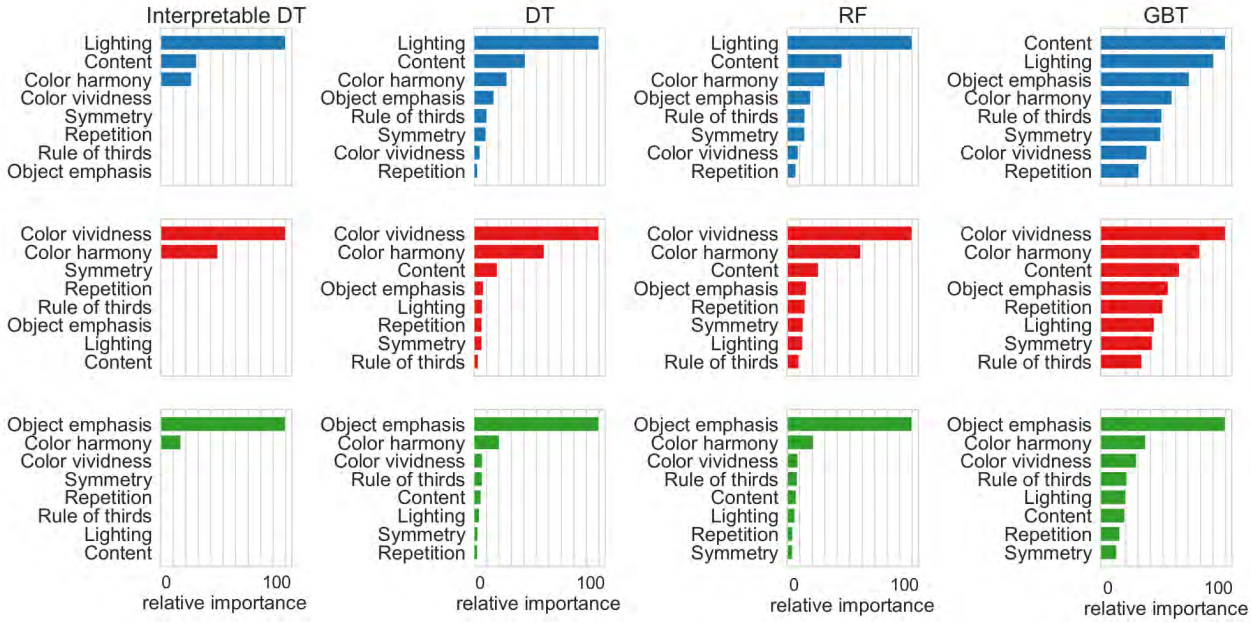

FIGURE 8. Relative feature importances for predicting aesthetics (top row), visual sentiment (middle row) and memorability (bottom row) scores for different regression models: CART based Decision Tree (DT) and its greedy pruning strategy variant (Interpretable DT), Random Forests (RF) and Gradient Boosting Trees (GBT).

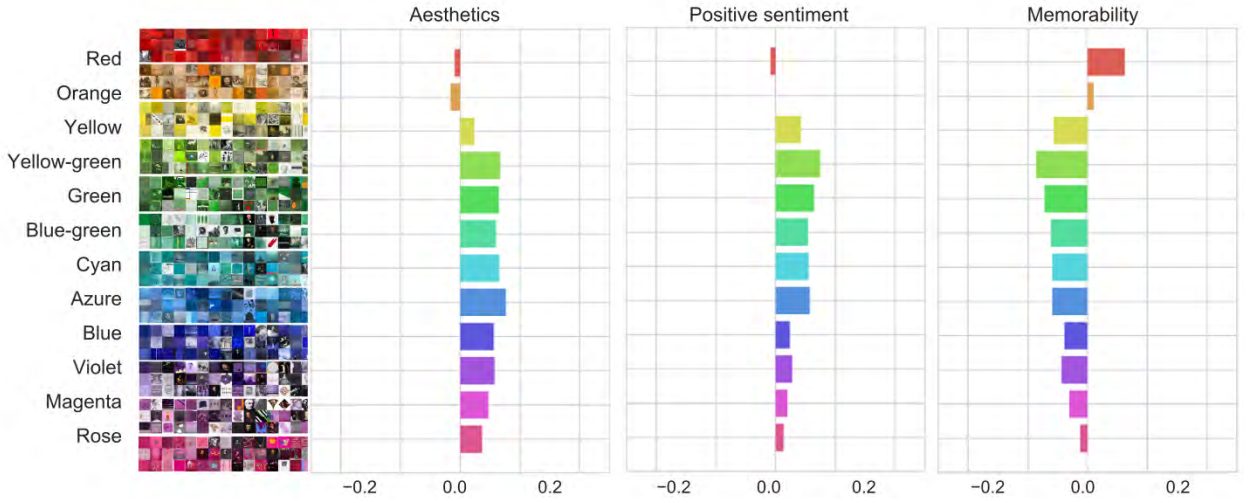

FIGURE 9. Spearman's rank correlation coefficients for aesthetics, sentiment and memorability scores and the amount of hue values obtained from a 12-bin normalized hue image histogram (p-values < 0.01).

following categories of interest: artistic style, genre, artists and chronology.

1) STYLE

In the context of the WikiArt dataset, the term "style" refers to a set of visual characteristics that are specific for a particular artistic movement, usually active in a certain time period. In order to explore the relation of aesthetic, sentiment and memorability features with different styles, we use a subset of 25 distinctive style categories that include more than 800 paintings. We calculate the mean aesthetic, sentiment and memorability scores for each style and show the distribution of predicted scores with the box plots presented in Fig. 10. The boxes are ordered by mean aesthetic score, which is marked with a blue dot.

Romanticism and Magic Realism are predicted as the most aesthetically pleasing categories (0.63), while Minimalism is the lowest ranked style with an average score of 0.49. However, the mean aesthetic scores are similar across different styles and it is difficult to differentiate styles based on the aesthetic scores, as well as to draw meaningful conclusions about the relation of styles and the predicted

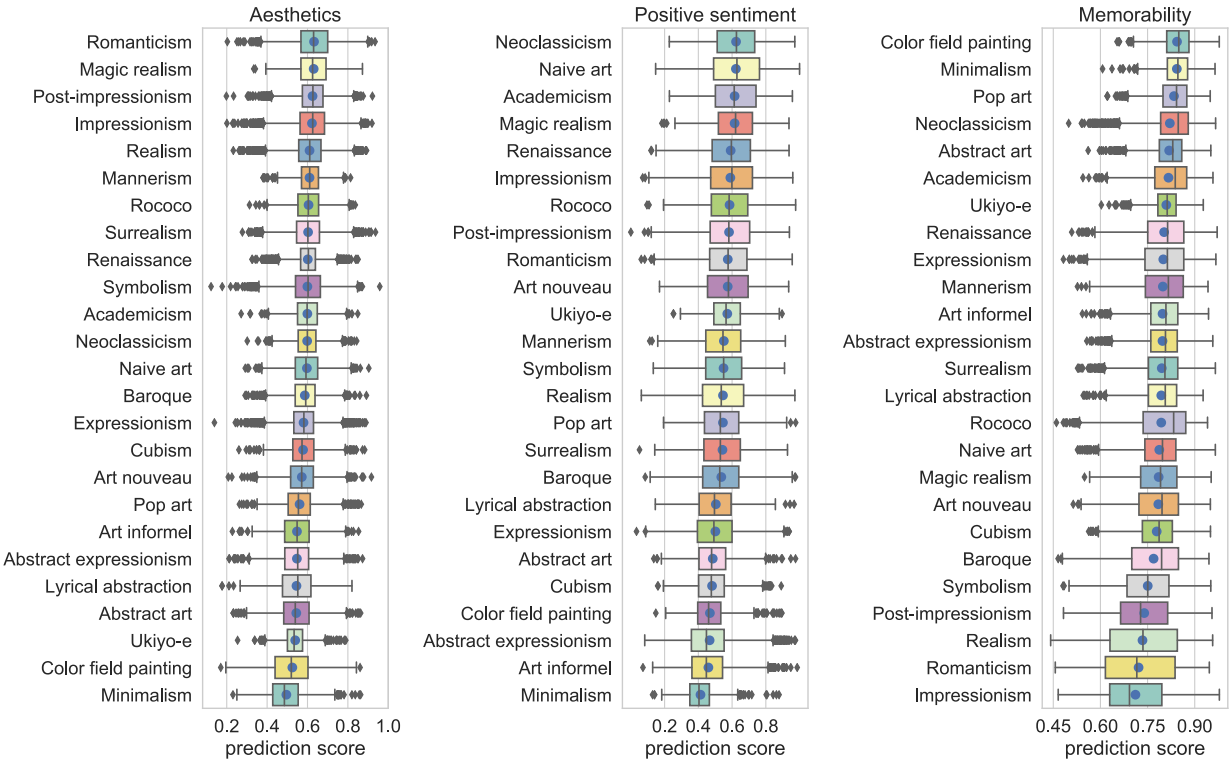

FIGURE 10. Box plot distribution of the image aesthetics, positive sentiment and memorability scores across artistic styles.

aesthetic scores. The distribution and mean values of the predicted positive sentiment scores are slightly more distinctive regarding different styles. Minimalism also has the lowest average sentiment score (0.39), while Neoclassicism, Naive art and Academicism have high average sentiment scores. The low memorability of Impressionism and Romanticism can be linked to the fact that those styles predominantly include landscape paintings. The high memorability score of abstract styles indicates that the absence of recognizable object content contributes to the increase of image memorability. This might be because visual stimuli in abstract paintings rarely appear in our daily visual experiences and therefore represent an exception that draws the viewer's attention.

2) GENRE

The term ''genre'' refers to the traditional division of paintings based on the type of content depicted, such as landscape, portrait, still life, etc. The WikiArt dataset includes a broad set of genre annotations and we focus on genre categories which correspond to specific objects or types of scenes. We select 13 different genre categories and, similarly as in the case of style categories, calculate the mean aesthetic, positive sentiment and memorability scores for each genre. The box plots in Fig. 11 show the distribution of the predicted scores across different genres.

The genre-specific mean aesthetic scores show that landscapes and cityscapes have the highest average aesthetic score, while abstract paintings have the lowest score. Although the difference between different genre-specific mean aesthetic scores is too small to make a significant distinction of genres in relation to aesthetic scores, it is interesting to notice how the predicted scores relate to content. Namely, paintings that include motifs related to nature (e.g. landscape, sea, animals) tend to have a higher aesthetic score than abstract paintings. Our results are consistent with some existing studies regarding aesthetic preferences in art. In particular, a broad and cross-cultural survey of art preferences among people presented in [3] suggests that people generally prefer figurative over abstract paintings, as well as motifs such as water, plants and animals. Genre-dependent memorability scores show that nude paintings have the highest average memorability, while the lowest score is obtained for landscapes. This is consistent with the finding that pictures with people tend to be more memorable than natural landscapes, presented by Isola *et al.* [43]. Because nude paintings and portraits have the highest average memorability score, while landscape and marina paintings have the lowest score, we might presume a consistency between memorability of art images and photographs when the subject of depiction is considered.

The flower painting category has also the highest mean positive prediction score, together with landscapes and animal paintings, while abstract paintings, cityscapes and battle paintings have a low average sentiment score. This result corresponds to the visual properties of the images with the

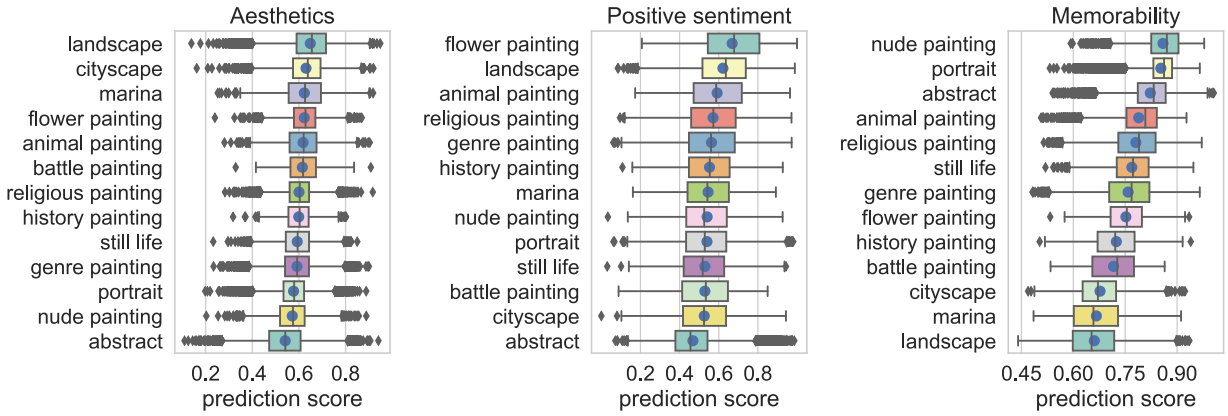

FIGURE 11. Box plot distribution of the image aesthetics, positive sentiment and memorability scores across artistic genres.

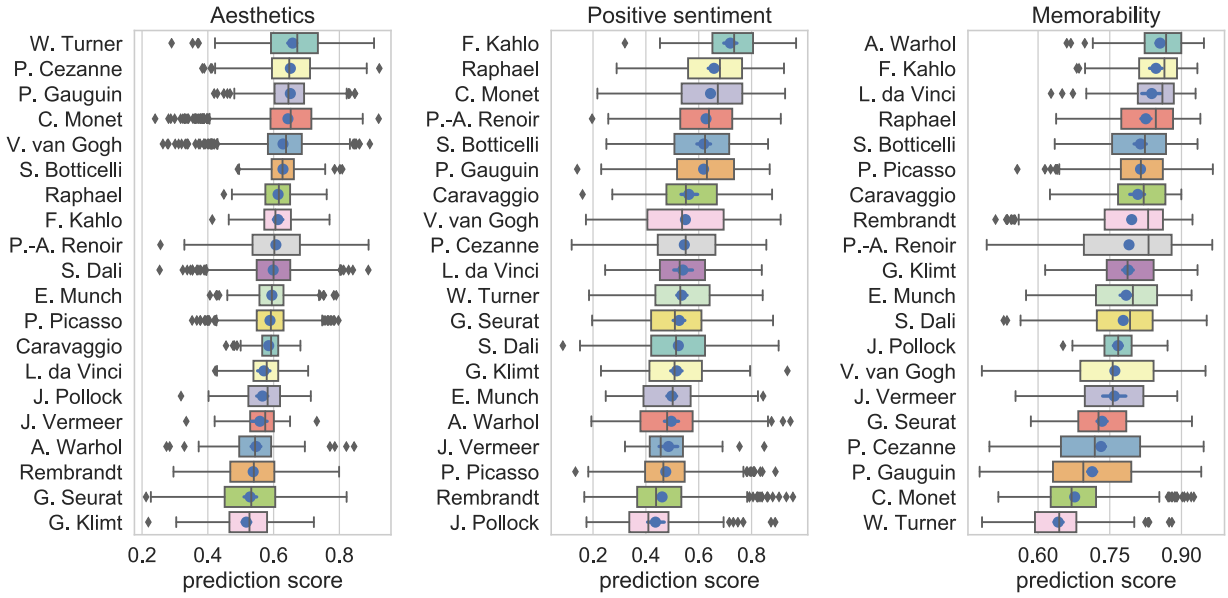

FIGURE 12. Box plot distribution of the image aesthetics, positive sentiment, and memorability scores across different artists.

highest and lowest positive sentiment scores shown in Fig. 5. The fact that battle paintings have a low average positive sentiment score, indicates a strong semantic correspondence with human judgment.

3) ARTISTS

The WikiArt dataset includes artworks by more than 2000 different artist, represented with a varying number of images. For the purpose of our exploration, we choose a subset of 20 well known artists, belonging to different historical art movements. Box plots in Fig. 12 show the distribution of the predicted scores for different artists.

The arbitrary choice of artists hinders us from making any general conclusions, however the relative relations between the predicted aesthetic, sentiment and memorability scores of the chosen artists still yield interesting outcomes. For instance, the case of William Turner whose works have the highest average aesthetic score and the lowest memorability score, prompts us to better understand how specific attributes of his works contribute to predicting low scores. As his work primarily consists of landscapes and marine paintings, this could explain the low memorability score. Interestingly, in a study which reports crowdsourced aesthetic ratings of artworks by different artists [63], Turner is also listed as having the highest average aesthetic score. In order to validate our aesthetic scores, we compare the aesthetic ranking order of six different artists, with a large number of estimates reported in the aforementioned study, with the average predicted aesthetic scores of images in the WikiArt dataset belonging to the same six artists. In the Supplemental files (Fig. S10) we provide a visual comparison which shows that the ranking order of artists based on the predicted scores is similar to the ranking order based on the crowdsourced aesthetic ratings.

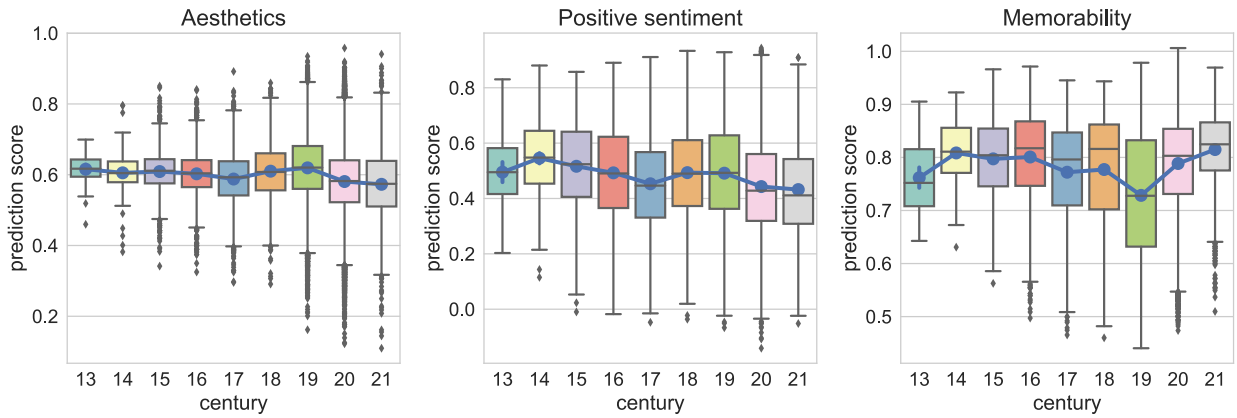

FIGURE 13. Box plot distribution of the image aesthetics, positive sentiment, and memorability scores across centuries.

An intriguing outcome is that the convincingly highest positive sentiment score, as well as a high memorability score, is obtained for the only female artist in this group - Frida Kahlo. By observing the artists who appear to have high sentiment and memorability scores, it is obvious that the vibrancy and richness of color plays an important role in predicting high scores. However, an interesting direction of future research is to investigate in more detail how characteristics of artistic oeuvres, in particular how the specific combination of subject and form, which creates distinctive individual artistic expressions, relates to the sentiment and memorability of the image.

4) CHRONOLOGY

Chronological ordering of information enables us to better understand the behavior of specific phenomena and to more easily identify interesting patterns. The WikiArt dataset contains information about the artwork's year of creation, although it is not available for all the artworks in the dataset. We use a subset of 82000 images for which the year of creation is known and group them by century, starting from the 13th to the 21st. Fig. 13 shows the mean values and the distribution of the predicted scores across the centuries.

We can see how the chronological curves of the predicted aesthetic and positive sentiment scores show similar behavior, with both reaching lower points in the 17th, 20th and 21st centuries. For both aesthetics and positive sentiment, the artworks from the 19th century tend to have high average aesthetic and sentiment scores, although the positive sentiment reaches its peak in the 14th century. Similarly, memorability scores tend to be high for the 14th and 21st century, but the lowest point is reached in the 19th century. This corresponds to the analysis of different styles, where the lowest average memorability score is obtained for styles belonging to the 19th century such as Impressionism, Romanticism and Realism.

VI. CONCLUSION

In this article we introduce a novel approach for a large-scale analysis of high-level features on art images. We investigate three perceptually relevant properties: the aesthetic evaluation of the image, the sentiment induced by the image and the memorability of the image. We use CNNs pre-trained for predicting these properties in natural images in order to obtain aesthetic, sentiment and memorability scores of art images. We analyze the relation of the obtained scores with other high-level image properties, as well as art history related concepts of style, genre, artists and chronology.

An interesting outcome of this analysis is that abstract styles tend to be more memorable, but have a lower average aesthetic and positive sentiment score. Furthermore, the genre-based distribution of scores, where the content of depiction plays the most important role, corresponds to previous photography-related findings and demonstrates consistency between art and natural images, as well as compatibility with intuitive human presumptions. The prediction of aesthetic, sentiment and memorability evaluation is particularly questionable in relation to individual artists. Even if a consensus regarding the relation of specific visual properties and the notion of beauty, sentiment and remembrance of images would exist, reducing the aesthetic evaluation of artists only to this relation is limited because it neglects their position in a broader historical and social context. However, in the context of computational image analysis, artistic oeuvres have a rare quality of producing image subsets with a sophisticated level of uniqueness and therefore represent a fruitful data source for advancing computationally developed image features.

Our qualitative and quantitative results suggest that CNN models pre-trained on natural images can extract meaningful aesthetic, sentiment and memorability features in the domain of fine art images. However, limitations emerge based on the choice of a particular task-specific model. Although the results obtained from differently trained models are consistent for the task of memorability, the aesthetic predictions are less consistent and depend more stongly on the choice of dataset and model architecture.

The confirmation of conclusions emerging from our results requires ground-truth labeling of the considered image properties on the same dataset, which requires complex experimental surveys. Nevertheless, the scores predicted using

CNNs represent an interesting finding and can serve as a basis for formulating initial hypotheses regarding the physiological relation of different high-level image properties. In the context of art history, the methodology presented in this article outlines novel directions for future research in computational analysis of artworks and domain-specific knowledge discovery. Knowing that the importance of a particular artwork does not only emerge from its visual properties, but also highly depends on the historical context, we are aware of the limitations of the proposed approach. However, the focus of this study is primarily oriented towards connecting the traditional formal analysis of art with computer vision and machine learning methods. In our future research we aim to transition the applicability of CNN in the context of art history to a more fine-grained level and address specific use cases relevant for art history-related research topics.

REFERENCES

[1] S. E. Palmer, K. B. Schloss, and J. Sammartino, "Visual aesthetics and human preference," *Annu. Rev. Psychol.*, vol. 64, pp. 77–107, Jan. 2013.

[2] G. M. Huebner and K. R. Gegenfurtner, "Conceptual and visual features contribute to visual memory for natural images," *PLoS One*, vol. 7, no. 6, Jun. 2012, Art. no. e37575.

[3] V. Komar and A. Melamid, *Painting by Numbers: Komar and Melamid's Scientific Guide to Art*. Berkeley, CA, USA: Univ California Press, 1999.

[4] R. N. Shepard, "Recognition memory for words, sentences, and pictures," *J. Verbal Learn. Verbal Behav.*, vol. 6, no. 1, pp. 156–163, Feb. 1967.

[5] S. Barnet, *A Short Guide to Writing About Art*. London, U.K.: Pearson, 2011.

[6] D. Keren, "Painter identification using local features and naive Bayes," in *Proc. Object Recognit. Supported Interact. Service Robots*, Quebec City, Canada, vol. 2, Aug. 2002, pp. 474–477.

[7] E. Cetinic and S. Grgic, "Automated painter recognition based on image feature extraction," in *Proc. 55th Int. Symp. ELMAR*, Sep. 2013, pp. 19–22.

[8] L. Shamir and J. A. Tarakhovsky, "Computer analysis of art," *J. Comput. Cult. Herit.*, vol. 5, no. 2, Jul. 2012, Art. no. 7.

[9] S. Agarwal, H. Karnick, N. Pant, and U. Patel, "Genre and style based painting classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 588–594.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[11] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., Sep. 2014, pp. 1–5.

[12] O. E. David and N. S. Netanyahu, "Deeppainter: Painter classification using deep convolutional autoencoders," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*. Barcelona, Spain: Springer, Sep. 2016, pp. 20–28.

[13] Y. Bar, N. Levy, and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, Sep. 2014, pp. 71–84.

[14] E. Cetinic and S. Grgic, "Genre classification of paintings," in *Proc. Int. Symp. (ELMAR)*, Sep. 2016, pp. 201–204.

[15] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[16] N. van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognit.*, vol. 61, pp. 583–592, Jan. 2017.

[17] A. Lecoutre and B. Negrevergne, and F. Yger, "Recognizing art style automatically in painting with deep learning," in *Proc. 9th Asian Conf. Mach. Learn. (ACML)*, Seoul, South Korea, Nov. 2017, pp. 327–342.

[18] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," *Expert Syst. Appl.*, vol. 114 pp. 107–118, Dec. 2018.

[19] C. Sandoval, E. Pirogova, and M. Lech, "Two-stage deep learning approach to the classification of fine-art paintings," *IEEE Access*, vol. 7, pp. 41770–41781, 2019.

[20] B. Seguin, C. Striolo, I. diLenardo, and F. Kaplan, "Visual link retrieval in a database of paintings," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 753–767.

[21] E. J. Crowley and A. Zisserman, "In search of art," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, Sep. 2014, pp. 54–70.

[22] G. Strezoski and M. Worring, "Omniart: A large-scale artistic benchmark," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, Nov. 2018, Art. no. 88.

[23] G. Gando, T. Yamada, H. Sato, S. Oyama, and M. Kurihara, "Fine-tuning deep convolutional neural networks for distinguishing illustrations from photographs," *Expert Syst. Appl.*, vol. 66, pp. 295–301, Dec. 2016.

[24] A. Elgammal, B. Liu, D. Kim, M. Elhoseiny, and M. Mazzone, "The shape of art history in the eyes of the machine," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, Feb. 2018, pp. 2183–2191.

[25] A. Brachmann, E. Barth, and C. Redies, "Using cnn features to better understand what makes visual artworks special," *Frontiers Psychol.*, vol. 8, p. 830, May 2017.

[26] G. U. Hayn-Leichsenring, T. Lehmann, and C. Redies, "Subjective ratings of beauty and aesthetics: Correlations with statistical image properties in western oil paintings," *I-Perception*, vol. 8, no. 3, pp. 1–21, May/Jun. 2017.

[27] S. Xu, H. Jiang, F. C. Lau, and Y. Pan, "Computationally evaluating and reproducing the beauty of chinese calligraphy," *IEEE Intell. Syst.*, vol. 27, no. 3, pp. 63–72, May 2012.

[28] R. Sun, Z. Lian, Y. Tang, and J. Xiao, "Aesthetic visual quality evaluation of chinese handwritings," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, Jun. 2015, pp. 2510–2516.

[29] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 1784–1791.

[30] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 1657–1664.

[31] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2408–2415.

[32] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, Sydney, NSW, Australia, Jan. 2015, pp. 524–535.

[33] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, New York, NY, USA, Nov. 2014, pp. 457–466.

[34] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 662–679.

[35] X. Jin, L. Wu, X. Zhang, J. Chi, S. Peng, S. Ge, G. Zhao, and S. Li, "Ilgnet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Comput. Vis.*, vol. 13, no. 2, pp. 206–212, Mar. 2018.

[36] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017.

[37] A. Brachmann and C. Redies, "Computational and experimental approaches to visual aesthetics," *Front. Comput. Neurosci.*, vol. 11, p. 102, Nov. 2017.

[38] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2010, pp. 83–92.

[39] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2013, pp. 223–232.

[40] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li, "Visual sentiment prediction with deep convolutional neural networks," 2014, *arXiv:1411.5731*. [Online]. Available: https://arxiv.org/abs/1411.5731

[41] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, Texas, USA, Jan. 2015, pp. 381–388.

[42] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, Sep. 2017.
[43] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1469–1482, Jul. 2014.
[44] M. Mancas and O. L. Meur, "Memorability of natural scenes: The role of attention," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, VIC, Australia, Sep. 2013, pp. 196–200.
[45] L. Goetschalckx, S. Vanmarcke, P. Moors, and J. Wagemans, "Are memorable images easier to categorize rapidly?" *J. Vis.*, vol. 17, no. 10, p. 98, 2017.
[46] X. Amengual, A. Bosch, and J. L. de la Rosa, "How to measure memorability and social interestingness of images: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 2, Feb. 2017, Art. no. 1754004.
[47] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2390–2398.
[48] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "Amnet: Memorability estimation with attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6363–6372.
[49] *Jenaaesthetics Dataset*. Accessed: Oct. 10, 2018. [Online]. Available: http://www.inf-cv.uni-jena.de/en/jenaesthetics
[50] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe, "In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 349–358.
[51] S. Mohammad and S. Kiritchenko, "Wikiart emotions: An annotated dataset of emotions evoked by art," in *Proc. 11th Int. Conf. Language Resour. Eval. (LREC)*, 2018, pp. 1–14.
[52] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 638–647.
[53] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2837–2841.
[54] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proc. CVPR*, Jun. 2011, pp. 145–152.
[55] S. A. Amirshahi, J. Denzler, and C. Redies, "JenAesthetics—A public dataset of paintings for aesthetic research," Comput. Vis. Group, Univ. Jena, Jena, Germany, Tech. Rep., 2013.
[56] S. A. Amirshahi, C. Redies, and J. Denzler, "How self-similar are artworks at different levels of spatial resolution?" in *Proc. Symp. Comput. Aesthetics*, Jul. 2013, pp. 93–100.
[57] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies, "Jenaesthetics subjective dataset: Analyzing paintings by subjective scores," in *Proc. Comput. Vis.-Workshops (ECCV)*, Zürich, Switzerland, Sep. 2014, pp. 3–19.
[58] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[59] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.
[60] W. Loh, "Classification and regression trees," *Wiley Interdisc. Rew., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[61] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[62] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[63] I. Mandel. (2018). *Aesthetic, Art-Historical and Economic Values in Painting: Empirical Study*. [Online]. Available: https://ssrn.com/abstract=3160419

EVA CETINIC received the M.Sc. degree in information and communication technology from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2012. In the same year, she enrolled in the Ph.D. Program with the Faculty of Electrical Engineering and Computing, University of Zagreb. Since 2015, she has been a Professional Associate with the Centre for Informatics and Computing, Rudjer Boskovic Institute. Her research interests include computer vision, image processing, and machine learning and their applications to digital arts and humanities-related data.

TOMISLAV LIPIC received the Ph.D. degree in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb. He is currently a Research Associate with the Laboratory for Machine Learning and Knowledge Representation, Rudjer Boskovic Institute. He was a Visiting Research Scholar with the Center for Polymer Studies, Boston University, MA, USA. He has participated in more than a dozen different national, bilateral and EU (FP7 and Horizon2020) industries, and scientific projects and project initiatives in diverse topic areas focused around theory and application of data science. His main research interests include complex networks modeling and analysis, network representation learning, interpretable and scalable machine learning and deep learning, other data science and big data methodologies and their applications in computational social sciences, neuroscience, and biomedicine.

SONJA GRGIC received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 1989, 1992, and 1996, respectively, where she is currently a Professor in multimedia technologies and communication systems. Her research interests include image processing and machine learning, picture quality evaluation, video communication technologies, and image forensics.