The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

The Shape of Art History in the Eyes of the Machine

Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny
The Art and AI Laboratory, Department of Computer Science
Rutgers University, NJ, USA

Marian Mazzone
Department of Art History, College of Charleston, SC, USA

Abstract

How does the machine classify styles in art? And how does it relate to art historians' methods for analyzing style? Several studies showed the ability of the machine to learn and predict styles, such as Renaissance, Baroque, Impressionism, etc., from images of paintings. This implies that the machine can learn an internal representation encoding discriminative features through its visual analysis. However, such a representation is not necessarily interpretable. We conducted a comprehensive study of several of the state-of-the-art convolutional neural networks applied to the task of style classification on 67K images of paintings, and analyzed the learned representation through correlation analysis with concepts derived from art history. Surprisingly, the networks could place the works of art in a smooth temporal arrangement mainly based on learning style labels, without any a priori knowledge of time of creation, the historical time and context of styles, or relations between styles. The learned representations showed that there are a few underlying factors that explain the visual variations of style in art. Some of these factors were found to correlate with style patterns suggested by Heinrich Wölfflin (1846-1945). The learned representations also consistently highlighted certain artists as the extreme distinctive representative of their styles, which quantitatively confirms art historian observations.

Introduction

Style is central to the discipline of art history. The word "style" is used to refer to the individual way or manner that someone makes or does something, for example Rembrandt's style of painting. Style also refers to groups of works that have a similar typology of characteristics, such as the Impressionist style, or High Renaissance style. Art historians are obligated to identify, characterize, and define styles based on the evidence of the physical work itself, in combination with an analysis of the cultural and historical features of the time and place in which it was made. Although we see style, and we all know that it exists, there is still no central, agreed upon theory of how style comes about, or how and why it changes. Some of the best scholars of art history have written persuasively about the importance of style to the discipline, and the concomitant difficulty of defining or explaining what it is and why it changes (Schapiro and Kroeber 1953; Gombrich 1968). Up to now connoisseurship has proven to be one of the most effective means to detect the styles of various artists, and differentiate style categories and distinctions in larger movements and periods.

Recent research in Artificial Intelligence have shown the ability of the machine to learn to discriminate between different style categories such as Renaissance, Baroque, Cubism, etc., with reasonable accuracy, (e.g. (Shamir et al. 2010; Karayev et al. 2013; Saleh et al. 2016; Saleh and Elgammal 2015)). However, classifying style by the machine is not what interests art historians. Instead, the important issues are what machine learning may tell us about how the characteristics of style are identified, and the patterns or sequence of style changes. None of the aforementioned papers brought any understanding of the problem of style that can be useful to art historian. The ability of the machine to classify styles implies that the machine has learned an internal representation that encodes discriminative features through its visual analysis of the paintings. However, it is typical that the machine uses visual features that are not interpretable by humans. This limits the ability to discover knowledge out of these results.

Our study's emphasis is on understanding how the machine achieves classification of style, what internal representation it uses to achieve this task, and how that representation is related to art history methodologies for identifying styles. To achieve such understanding, we utilized one of the key formulations of style pattern and style change in art history, the theory of Heinrich Wölfflin (1846-1945). Wölfflin's comparative approach to formal analysis has become a standard method of art history pedagogy. Wölfflin chose to separate form analysis from discussions of subject matter and expression, focusing on the "visual schema" of the works, and how the "visible world crystallized for the eye in certain forms" (Wölflin 1950). Wölfflin identified pairs of works of art to demonstrate style differences through comparison and contrast exercises that focused on key principles or features. Wölfflin used his method to differentiate the Renaissance from the Baroque style through five key visual principles: linear/painterly, planar/recessional, closed form/open form, multiplicity/unity, absolute clarity/relative clarity. Wölfflin posited that form change has some pattern of differentiation, such that style types and changes can only come into being in certain sequences.

With advances in AI and the availability of comprehensive datasets of images, we are now positioned to approach the history of art as a predictive science, and relate its means of determining questions of style to machine results. It was nearly impossible to apply and empirically test Wölfflin's methods of style differentiation and analysis before developments in AI. No human being would be able to assemble or analyze the number of examples needed to prove the value of his methods for finding discriminative features.

Methodology: Deep convolutional neural networks have recently played a transformative role in advancing artificial intelligence (LeCun, Bengio, and Hinton 2015). We evaluated a large number of state-of-the-art deep convolutional neural network models, and variants of them, trained to classify styles. We focused on increasing the interpretability of the learned presentation by forcing the machine to achieve classification with a reduced number of variables without sacrificing classification accuracy. We then analyzed the achieved representations through linear and nonlinear dimensionality reduction of the activation space, visualization, and correlation analysis with time and with Wölfflin's pairs.

Detailed Methodology

Challenges with Art Style Classification: In contrast to the typical object classification in images, the problem of style classification in art is different and has different set of challenges. Style is not necessarily correlated with subject matter, which corresponds to existence of certain objects in the painting. Style is mainly related to the form and can be correlated with features at different levels, low, medium and high-level. As a result, it is not necessary that networks which perform better for extracting semantic concepts, such as object categories would be perform as well in style classification. In the literature of object classification, deeper networks were shown to perform better (Simonyan and Zisserman 2014; He et al. 2016) since it facilitates richer representation to be learned at different levels of features. We do not know if a deeper network will necessarily be better in the style classification domain. This remains something to discover through our empirical study. However, the challenge, in the context of art style classification, is the lack of images on a scale of magnitude similar to ImageNet (million of images). The largest publicly available dataset, which we use, is only in the order of 70K images. This limitation is due to copyright issue, which is integral to the domain of art. Moreover, collecting annotation in this domain is hard since it requires expert annotators and typical crowd sourcing annotators are not qualified.

Another fundamental difference is that styles do no lend themselves to discrete mutually exclusive classes as in supervised machine learning. Style transition over time is typically smooth, and style labels are after-the-fact concepts imposed by art historians, sometimes centuries later. Paintings can have elements that belongs to multiple styles, and therefore not necessarily identifiable with a unique style class.

Datasets: *Training-testing Set:* We trained, validated, and tested the networks using paintings from the publicly available WikiArt dataset[1]. This collection (as downloaded in 2015) has images of 81,449 paintings from 1,119 artists ranging from the fifteenth century to contemporary artists. Several prior studies on style classification used subsets of this dataset (e.g. (Karayev et al. 2013; Saleh et al. 2016; Saleh and Elgammal 2015)). For the purpose of our study we reduced the number of classes to 20 classes by merging fine-grained style classes with small number of images[2]. We excluded from the collections images of sculptures and photography. The total number of images used for training, validation, and testing are 76,921 images. We split the data into training (85%), validation (9.5%) and test sets (5.5%).

Visualization Set I: We used another smaller dataset containing 1485 images of paintings from the Artchive dataset[3] to analyze and visualize the representation. Previous researches that have used this dataset (e.g. (Saleh et al. 2016)). While the WikiArt collection is much bigger in size, this dataset contains a better representation of the important works of western art from 1400-2000AD by 60 artists. Therefore, we mainly use it to visualize and analyze the learned representation.

Visualization Set II: We also used 62K painting from the Wikiart dataset for visualization and analysis of the representations. We only included paintings that have date annotation for the purpose of visualization and time correlation studies.

Wölfflin's pairs' annotation: We also collected art historian's rating annotations (scale of 1 to 5) for each of the Wölfflin's pairs for 1000 paintings from this data set and use it in our correlation analysis.

Studied Deep Learning Models: We performed a comparative study on three networks: AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGGNet (Simonyan and Zisserman 2014), and ResNet (He et al. 2016), as well as variants of them, adapted for the task of style classification. All these models were originally developed for the task of object recognition for the ImageNet challenge (Russakovsky et al. 2015) and each of them raised the state of the art when they were introduced. For all the models, the final softmax layer, originally designed for the 1000 classes in ImageNet, was removed and replaced with a layer of 20 softmax nodes, one for each style class. Our study included varying the training strategies (training from scratch on art data vs. using pre-trained models and fine-tuning them on art data), varying the network architecture, and data augmentation strategies. Fine-tuning is the standard practice when adapting well-performing pre-trained models to a different domain.

Increasing the interpretability of the representation:

Having a large number of nodes at the fully connected layer allows the representation to project the data into a very high dimensional space where classification would be easy (specially in our case with only 20 classes), without enforcing similar paintings across styles to come closer in the representation. To increase the interpretability of the

[1] Wikiart dataset http://www.wikiart.org

[2] The full list of style classes and merges is available in the supplementary materials Table 1

[3] Artchive dataset http://www.artchive.com

representation, we force the network to achieve classification through a lower dimension representation. To achieve this, after training the network (whether from scratch or through fine-tuning), two more fully connected layers were added with a reduced number of nodes. These reduced dimensional layers enforce the representation to use a smaller number of degrees of freedom, which in turn enforces paintings across styles to come closer in the representation based on their similarity. In particular, we added two layers with 1024 and 512 nodes to all the models, and the models are then fine-tuned to adjust the weights for the new layers. As will be shown later, adding these dimensionality reduction layers did not affect the classification accuracy. The experiments showed that gradually reducing the number of nodes in the fully connected layers forces the network to achieve a "smoother" and interpretable representation.

Quantitative Comparative Results

Style Classification - Comparative Results Table 1 shows the classification accuracy of different models using both pre-training with fine-tuning, and training from scratch. In all cases, the pre-trained and fine-tuned networks achieved significantly better results than their counterparts that are trained from scratch (7% to 18% increase). This is not surprising and consistent with several models that adapted and fine-tuned pre-trained networks for different domains. However, learned filters when the network trained from scratch on the art domain were significantly different from the ones typically learned on ImageNet, which typically shows Gabor-like and blob-like filters (see Figure SM2). While it is hard to interpret the filters trained for style classification, we do not observe oriented-edge-like filters except of a horizontal edge filter. This emphasizes the different in nature between the problems and suggests that the better performance of the fine-tuned models could be out-performed if sufficient data is available to train a style-classification network from scratch on art data only.

Table 1: Comparison of classification of different models and different training methodologies

Network	**Architecture**	**Trained** from scratch	**Pre-trained & Fine-tuned**
AlexNet	Original:5 Conv layers+3 FC layers	48.1%	58.2%
AlexNet+2	Adding 2 reduced FC layers	47.2%	58.3%
VGGNet	13 Conv layers + 3 FC layers	51.6 %	60.1%
VGGNet+2	Adding 2 reduced FC layers	55.2%	62.8%
ResNet	152 Conv layers		63.7%
	50 Conv layers	45.0%	
ResNet+2	152 Conv layers +2 reduced FC layers		60.2%
	50 Conv layers + 2 reduced FC layers	48.0%	

Increasing the depth of the network only added no more than 5% to the accuracy from AlexNet with 5 convolutional layers to ResNet with 152 convolutional layer. In the case for learning from scratch, increasing the depth didn't improve the results where a ResNet with 50 layers performed worse than an AlexNet with only 5 convolution layers. VGGNet with 13 convolutional layers performed only 2-3% better than AlexNet. Increasing the depth of VGGNet didn't improve the results. This limited gain in performance with increase in depth, in conjunction with the difference in the learned filters, suggests that a shallow network might be sufficient for style classification along with better filter design.

Table 2: The effect of adding two reduced dimension layers on the representation (subspace dimensionality and variance)

Model	Training Strategy	Original Architecture			Adding two dimensionality reduction layers		
		number of nodes	subspace dim [1]	retained variance [2]	number of nodes	subspace dim[1]	retained variance[2]
AlexNet	Pre-trained & Finetuned	4096	201	21.71	512	9	59.64
AlexNet	From Scratch	4096	397	35.62	512	10	62.78
VGGNet	Pre-trained & Finetuned	4096	55	49.52	512	7	66.87
VGGNet	From Scratch	4096	36	51.16	512	7	72.52
ResNet	Pre-trained & Finetuned	2048[3]	491	17.53	512	6	73.71

[1] Subspace dim: Number of principle components cumulatively retaining 95% of variance.
[2] Retained variance: Percentage of variance retained by the first two principle components.
[3] ResNet does not have FC layers. This is the number of the nodes in the last pooling layer.

Effect of adding layers with reduced dimensionality The experiments showed that adding extra fully connected layers with gradually reducing the number of nodes in them forces the networks to achieve a "smoother" and more interpretable representation. We quantified this phenomenon by examining the dimensionality of the subspace of the activation (using Principle Component Analysis (PCA (Jolliffe 2002)) of the visualization dataset using two measures: 1) The number of components needed to preserve 95% of the variance. 2) The variance retained with the first two PCA dimensions. We also evaluated the accuracy of the expanded models to see if the added reduced layers resulted in any loss of accuracy. In most of the cased the added layers enhanced the accuracy (see Table 1)

Table2 shows that adding two reduced-dimension layers effectively and consistently reduced the dimensionality of the subspace of the data while preserving the classification accuracy. The reduction is significant for AlexNet where 9 or 10 dimensions retained 95% of the variance compared to 201 and 397 dimensions for the cases of fine-tuned and learned from scratch networks respectively, with around 60% of the variance retained in the first two dimension. Interestingly, the representation achieved by VGGNets already has reduced dimension subspaces compared to the AlexNet and ResNet. However, adding the reduced dimension layers for VGG significantly lowered the subspace dimension (only 7 dimensions retain 95% of the variance) while improving its classification accuracy between 2-4%. The maximum reduction in subspace dimensionality was in ResNet where the dimension of the subspace retaining 95% of the variance was reduced from 491 to only 6 with 74% of the variance in the first two dimensions (Figure SM3).

Table 3: Temporal correlation with the first two PCA dimensions and the first LLE dimensions of the activation space in different models

		Pearson correlation coefficient with time			
model	training	1^{st} PCA dim	2^{nd}PCA dim	Radial	1^{st}LLE dim
AlexNet+2	Fine-tuned	0.4554	0.5545	0.6944	0.7101
	From scratch	-0.5797	0.2829	0.6697	0.6723
VGGNet+2	Fine-tuned	-0.2462	0.5316	0.5659	-0.4012
	From scratch	0.5781	0.3165	0.6239	-0.6532
ResNet+2	Fine-tuned	-0.6559	0.4579	0.7712	0.8130

Table 4: Correlation with Wölfflin's concepts. Pearson Correlation Coefficient of the first two PC dimensions and the first two LLE dimensions of the activation space and Wölfflin's concepts. The concepts with maximum correlation with each dimension is shown.

		Pearson correlation coefficient(absolute values) with Wölfflin's concepts			
Model	Training	1^{st} PCA dim Planar vs. Recession	2^{nd}PCA dim Linear vs. Painterly	1^{st}LLE dim Planar vs. Recession	2^{nd}LLE dim and Linear vs. Painterly
AlexNet+2	Fine-tuned	0.4946	0.3579	0.501	0.3216
	From scratch	0.5129	0.3272	0.4930	0.3111
VGGNet+2	Fine-tuned	0.3662	0.2638	0.4512	0.2646
	From scratch	0.4621	0.4000	0.4897	0.3174
ResNet+2	Fine-tuned	0.5314	0.4795	0.5251	0.4158

Interpretation of the Representation

This section focuses on analyzing, visualizing, and interpreting the activation space induced by the different networks after trained to classify style. We define the activation space of a given fully connected layer as the output of that layer prior to the rectified linear functions. In particular, in this paper, we show the analysis of activation of the last reduced dimension fully connected layer prior to the final classification layer, which consists of 512 nodes. We use the activations before the rectified linear functions in all the networks.

A Few Factors Explains the Characteristics of Styles: The learned representation by the machine shows that there are a few underlying factors that can explain the characteristics of different styles in art history. Using PCA, we find that only fewer than 10 modes of variations can explain over 95% of the variance in the visualization set in all of the studied models with additional reduced fully connected layers. In ResNet and VGGNet the number of these modes is as low as 6 and 7 respectively. In all of the networks, the first two modes of variations explained from 60% to 74% from the variance in the various models in visualization set I (Table 1). Moreover, it is clear from the visualizations of both the linear and nonlinear embedding of the activation manifold in various models that art dated prior to 1900 lie on a plane (subspace of dimension 2).

Consistent results are achieved by analyzing the 62K painting from the Wikiart data set where it was found that subspaces of dimensions 10, 9 and 7 retain 95% of the variance of the activation for AlexNet+2, VGGNet+2, ResNet+2 respectively (see SM). The consistency of results in all the studied networks and the two datasets (varying in size from 1500 to 62K paintings) implies that the existence of a small number of the underlying factors explaining the representation is an intrinsic property of art history, and not just an artifact of the particular dataset or model.

We will start our interpretation of these dimensions by investigating the time correlation with these dimensions, followed by correlation analysis with Wölfflin's concepts.

Smooth Temporal Evolution of Styles: The results also indicate that the learned representations, projected to the first modes of variations of the activation space, show a smooth temporal transition between styles (Fig.1, SM4). This is despite the fact that the networks are trained only with images and their discrete style labels. No information was provided about when each painting was created, when each style took place, which artist created which painting, nor how styles are related (such as style x is similar to style y, or came after or before style z). Despite the lack of all these cues, the learned representations are clearly temporally smooth and reflect high level of correlation with time (Table 3).

Most interestingly, studying the modes of variations in the representation showed a radial temporal progress, with quantifiable correlation with time. In Figure 1-A, we can observe the temporal progress in a clock-wise way from Italian and Northern Renaissance at the bottom, to Baroque, to Neo-classicism, Romanticism, reaching to Impressionism at the top followed by Post- Impressionism, Expressionism and Cubism. A full cycle completes with 20th century styles such as abstraction and Pop Art coming back close to Renaissance. In fact, the angular coordinates have a Pearson Correlation Coefficient (PCC) of 0.69 with time. The strongest temporal correlation is in the case of ResNet with 0.76 PCC radial correlations. This conclusion is consistent among all the networks that we tested (Table 3). Our study shows that style changes smoothly over time and proves that noisy style labels are enough for the networks to recover a temporal arrangement, due mainly to visual similarities as encoded through the learned representation. This is consistent with Wölfflin's sequence hypothesis.

Relation to Wölfflin's Pairs: By studying the correlation between the modes of variations and Wölfflin's suggested pairs, we find that the first mode of variation in all the learned models consistently correlates the most with the concept of plane vs. recession, while the second mode of variation correlates the most with the concept of linear vs. painterly. This correlation explains the radial temporal progress and the loop closure between Renaissance and 20th century styles since they share linearity and planarity in their form. In Figure 1-A we can see the smooth transition from linear-planar form in Renaissance at the bottom towards more painterly-recessional form in Baroque to the extreme case of painterly with Impressionism at the top. Next we can see the transition back to linear-planar form in abstract and Pop art styles. Projecting the data in these two dominant modes of variations that are aligned with plane vs. recession and linear vs. painterly explains why this representation correlates with time in a radial fashion.

Figure 1-C shows the fourth and fifth dimensions of AlexNet+2 representation, which spreads away strongly the Renaissance vs. Baroque styles and put other styles in perspective to them. The fifth dimension (vertical) in particular correlates with absolute vs. relative clarity, multiplicity vs. unity, closed vs. open, and linear vs. painterly form from top to bottom (with PCC -0.36, -0.30, -0.28, -0.16 respectively). The fourth dimension correlates to a lesser degree with the opposite of these concepts. Therefore, in Figure 1-C the Renaissance style appears at the top (absolute clarity, multiplicity, closed, linear form), with the Baroque at the bottom (relative clarity, unity, and open, painterly form). This is quite consistent with Wölfflin's theory since he suggested exactly the same contrast between these concepts to highlight the difference between Renaissance and Baroque. We can also see in the figure that Impressionism and Cubism are toward the bottom of the plot since they share many of these same

concepts with Baroque. The fourth dimension seems to separate Impressionism and Cubism to the right from Abstract and Pop art to the left.

Identification of Representative Artists: Visualizing the different representations shows that certain artists were consistently picked by the machine as the distinctive representatives of their styles, as they were the extreme points along the dimensions aligned with each style. This is visible in the first three modes of variations of the representation learned by the VGGNet, shown in Figure 2A, which retains 77% of the variance (Consistent results on a large-scale collection is shown in Fig 2F-H). Besides the observable temporal progress, the representation separates certain styles and certain artists within each style distinctively from the cloud at the center as non-orthogonal axes. Through independent source separation[4], we factorized these axes, which are aligned with styles where extreme points on these axes represent distinguishable artists in each style (Fig2B-E, SM5). For example, such distinctively representative artists are Van Eyck and Dürer for Northern Renaissance, Raphael for Italian Renaissance, Rembrandt and Rubens for Baroque, Monet for Impressionism, Cézanne and Van Gogh for Post-Impressionism, Rousseau for Naïve-Primitivism, Picasso and Braque for Cubism, and Malevich and Kandinsky for abstract. The fact that the representation chose a certain representative artist or artists for each style – among thousands of paintings by many artists in each style – highlights quantitatively the significance of these particular artists in defining the styles they belong to.

Activation Manifold and Understanding Influences: We also studied the activation manifold of the different representation learned by the machine[5], which also reveals smooth temporal progress captured by the representation, as well as correlation with Wölfflin's concepts (Figure 3, SM Tables 3,4). Varying the neighborhood structure when constructing the manifold allowed us to discover different interesting connections in the history of art in a quantifiable way. For example, the manifold embedding in Figure 3-D shows how Cézanne's works are acting as a bridge between Impressionism at one side and Cubism and abstract art at the other side. Art historians consider Cézanne to be a key figure in the style transition towards Cubism and the development of abstraction in 20th century art. This bridge of Cézanne's paintings in the learned representation is quite interesting because that is a quantifiable connection in the data, not just a metaphorical term. Another interesting connection is that between the Renaissance and modern styles such as Expressionism, Abstract-Expressionism with the works of El-Greco, Dürer, Raphael, and other Renaissance painters (Figure 3-E). These artists are pulled away from the Renaissance cluster at the left as outliers to the temporal progress because of their similarity to several works in modern styles. Notably, both Cézanne's connection and the El-Greco/Dürer connection appear consistently in the various representations learned by different networks, however manifested in different forms.

Discovering Limitations of Wölfflin's Concepts: Interestingly, not all modes of variations explaining the data correlate with Wölfflin's concepts. In all learned representations, one of the first five modes of variation always has close-to-zero linear correlation with all Wölfflin's concepts. A notable example is the fifth dimension of the embedded activation manifolds, which separates Impressionism from Post-Impressionism, and has almost zero linear correlation with Wölfflin's pairs (Fig SM6). This implies that the separation between these styles is not interpretable in terms of Wölfflin's pairs.

Discussion and Conclusion

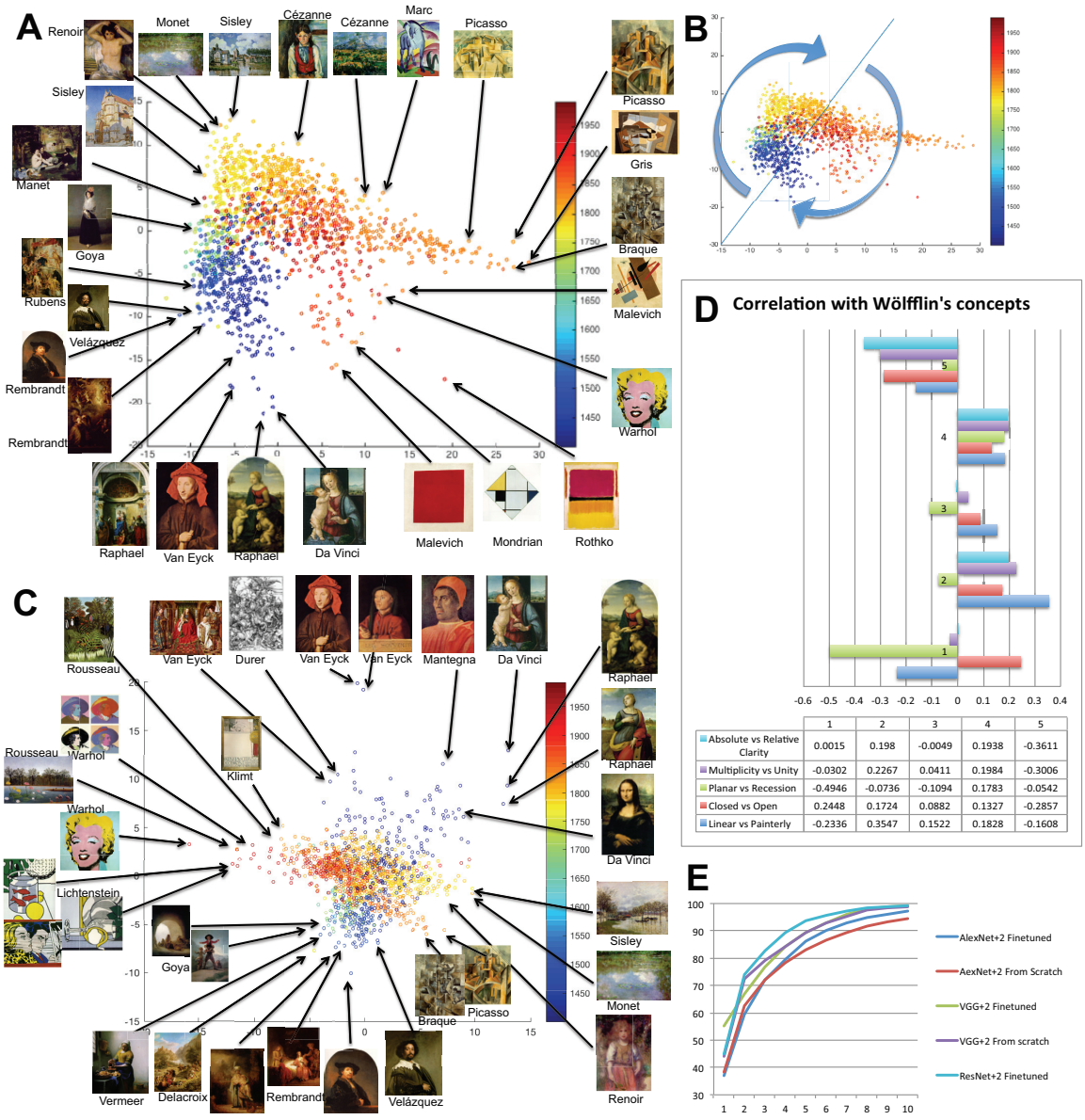
The consistency of the results among the different models and data sets indicate that the results are not just artifacts of the data or the algorithms, but rather are due to intrinsic properties of style change in art over history. The implication of the networks' ability to recover a smooth temporal progression through the history of art, in absence of any temporal cues given at training other than the constraint of putting paintings of the same style closer to each other to achieve classification, suggests that visual similarity is the main reason that forces this smooth temporal representation to evolve. This in turns echoes the smooth transition of style in the history of art. This is consistent with Wölfflin sequencing hypothesis. The analysis also indicates that a small number of factors encapsulate the visual characterization of different styles in art history.

The networks are presented by raw colored images, and therefore, they have the ability to learn whatever features suitable to discriminate between styles, which might include compositional features, contrast between light and dark, color composition, color contrast, detailed brush strokes, subject matter related concepts. In particular, networks pretrained on object categorization datasets might suggest potential bias towards choosing subject-matter-related features for classification. However, visualizing the learned representations reveals that the learned representations role out subject matter as a basis for discrimination. This is clear from noticing the loop closure between Renaissance style, which is dominated with religious subject matter, and modern 20th century styles, such as Abstract, Pop art, and others. In contrast, this loop closure suggests that the basis of discrimination is related to concepts related to the form as suggested by Wölfflin.

Overall the results highlight the potential role that AI can play in the domain of art history to discover patterns and trends. The study also highlights the importance of revisiting the formal methods in art history pioneered by art historians such as Wölfflin using tools from computer vision and machine learning.

[4]This is achieved using Independent Component Analysis using the FastICA algorithm (Hyvärinen, Karhunen, and Oja 2001; Hyvarinen 1999)

[5]We used Locally Linear Embedding (LLE) for achieving an embedding of the activation manifold. (Roweis and Saul 2000)



	1	2	3	4	5
Absolute vs Relative Clarity	0.0015	0.198	-0.0049	0.1938	-0.3611
Multiplicity vs Unity	-0.0302	0.2267	0.0411	0.1984	-0.3006
Planar vs Recession	-0.4946	-0.0736	-0.1094	0.1783	-0.0542
Closed vs Open	0.2448	0.1724	0.0882	0.1327	-0.2857
Linear vs Painterly	-0.2336	0.3547	0.1522	0.1828	-0.1608

Figure 1: Modes of variations of the activation subspace showing smooth temporal transition and correlation with Wölfflin's concepts. (A) First and second modes of variations of the AlexNet+2 model with paintings color-coded by date of creation. The first mode (the horizontal axis) seems to correlate with figurative art, which was dominant till Impressionism, vs. non-figurative, distorted figures, and abstract art that dominates 20th century styles. Another interpretation for this dimension is that it reflects Wölfflin's concept of plane (to the right) vs. recession (to the left). This axis correlates the most with Wölfflin's concept of plane vs. recession with -0.50 PCC. To a lesser degree, this horizontal axis correlates with closed vs. open (0.24 PCC) and linear vs. painterly (-0.23 PCC). This quantitative correlation can be clearly noticed by looking at the sampled paintings shown. The vertical axis correlates with the linear (towards the bottom) vs. painterly (towards the bottom) concept (0.36 PCC). (B) The angular coordinate exhibits strong correlation with time (PCC of 0.69). (C) The forth and fifth modes of variations show separation between Renaissance and Baroque as it correlates with Wölfflin's concepts distinguishing these styles. (D) Correlation of the first 5 modes of variations with Wölfflin's concepts. (E) The cumulative retained variance of the first 10 modes of variations of different models.

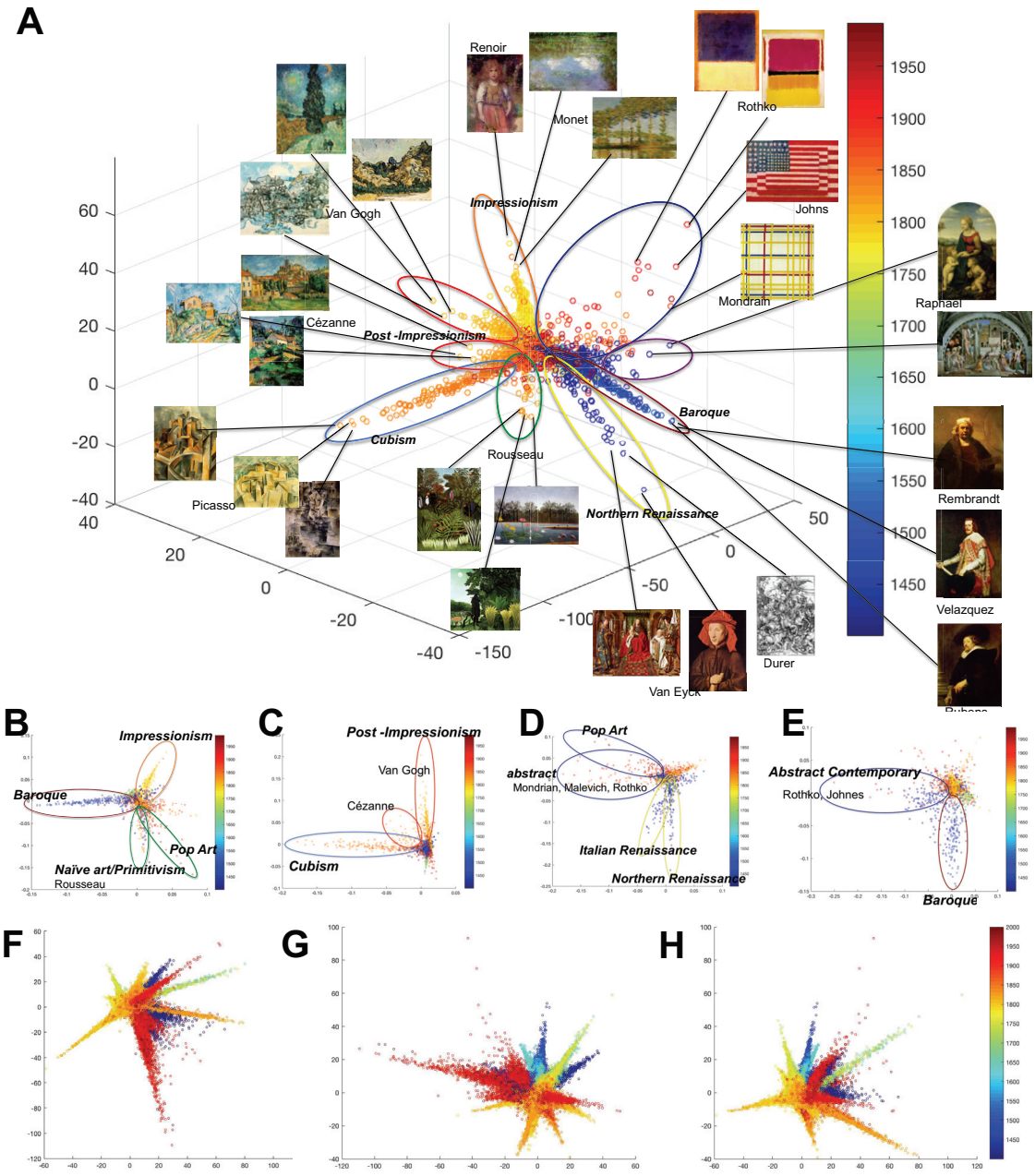


Figure 2: The representation could discover distinctive artists for each style. (A)The first three modes of variations of the VGGNet activations. Distinctive artists representing each style are identified by the representation and pulled away from the cloud at the center. We can see the Northern Renaissance in the yellow ellipse with the majority of the paintings sticking out being by Van Eyck and Dürer. The Baroque in the black ellipse is represented by Rubens, Rembrandt, and Velazquez. The orange ellipse is Impressionism and at its base are Pissarro, Caillebotte, and Manet as the least painterly of the type, ending with Monet and Renoir as most painterly on the end of the spike. The two red circles are Post-Impressionism, and in particular one is dominated by Van Gogh, and the other by Cézanne who forms the base for the spike of Cubism in the light blue ellipse. This spike is dominated by Picasso, Braque, and Gris; and goes out to the most abstract Cubist works. Most interestingly the representation separates Rousseau, as marked in the green ellipse, which is mainly dominated by his work. (B-E) Factorization of the activation space using Independent Component Analysis into 7 maximally independent axes, which show alignment with styles (more details in Fig S4, S5). (F-H) The top three modes of variations in the VGG network activation of 62K works of art from the Wikiart collection (projected pairwise as dimensions 1-2, 2-3, 3-1 from left to right).

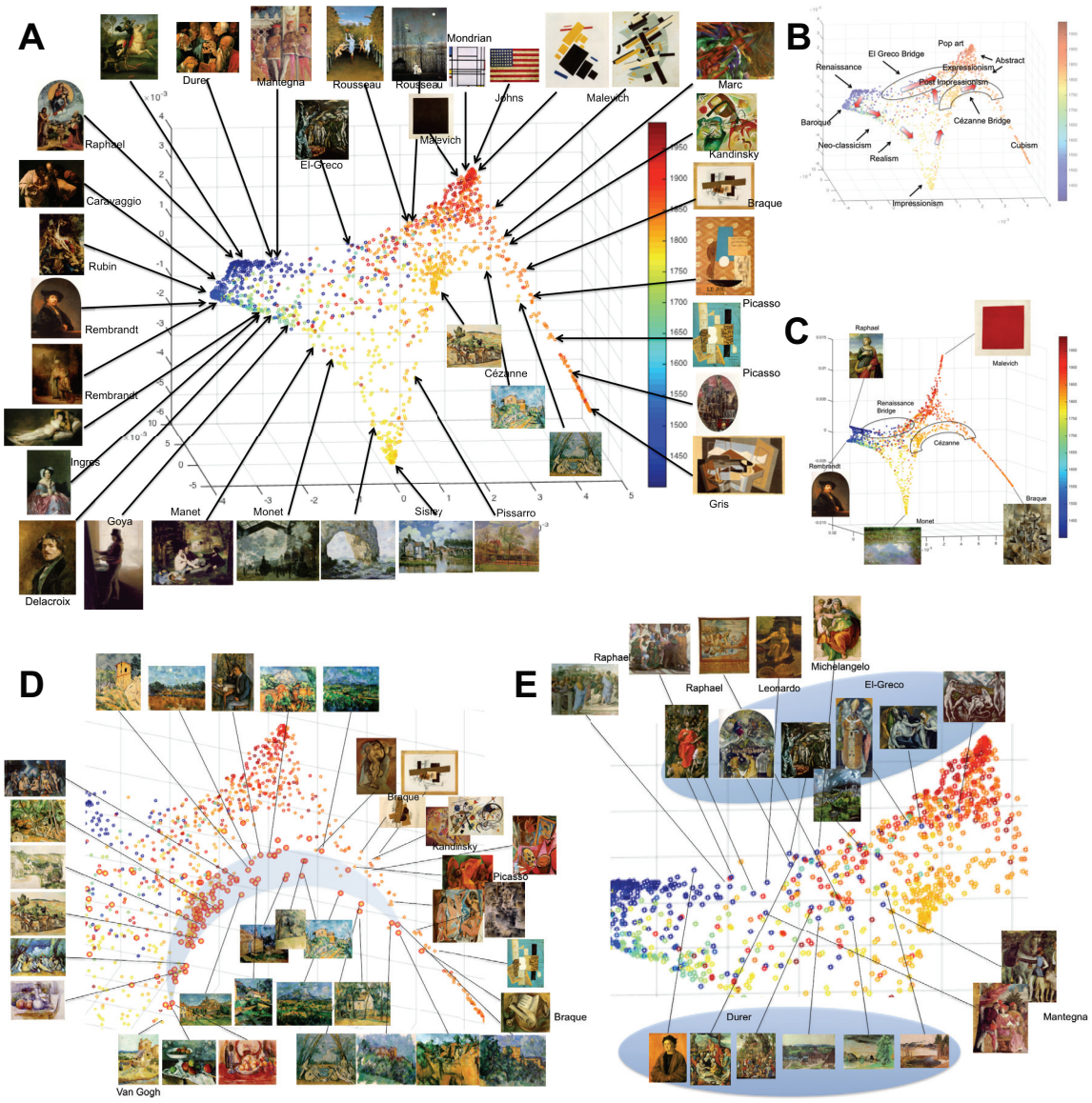


Figure 3: Interesting connections discovered in the activation manifold. (A) Example of activation manifold of AlexNet+2. Paintings color coded by date on the surface of the activation manifold showing smooth transition over time. (B) Transitions between art movements and important connections. (C) Accentuated version of the manifold highlighting five major styles: Renaissance, Baroque, Impressionism, Cubism, and abstract. (D) Cézanne's bridge: we can see branching at Post-Impressionism where Cézanne's work clearly separates from the other Post-Impressionist and Expressionist works towards the top. This branch continues to evolve till it connects to early Cubist works by Picasso and Braque, as well as abstracts works by Kandinsky. All thumbnails without labels in this plot are by Cézanne. (E) The connection between Renaissance and modern styles is marked by the outliers in the temporal progress patterns by certain works by El-Greco, Dürer, Raphael, Mantegna, and Michelangelo.

References

Gombrich, E. 1968. Style, international encyclopedia of the social sciences.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hyvärinen, A.; Karhunen, J.; and Oja, E. 2001. Independent component analysis. series on adaptive and learning systems for signal processing, communications, and control.

Hyvarinen, A. 1999. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks* 10(3):626–634.

Jolliffe, I. T. 2002. Principal component analysis and factor analysis. *Principal component analysis* 150–166.

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290(5500):2323–2326.

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. *International Journal of Computer Vision* 115(3):211–252.

Saleh, B., and Elgammal, A. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Saleh, B.; Abe, K.; Arora, R. S.; and Elgammal, A. 2016. Toward automated discovery of artistic influence. *Multimedia Tools and Applications* 75(7):3565–3591.

Schapiro, M., and Kroeber, A. L. 1953. Style. *Anthropology Today*.

Shamir, L.; Macura, T.; Orlov, N.; Eckley, D. M.; and Goldberg, I. G. 2010. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)* 7(2):8.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wölflin, H. 1950. *Principles of Art History: The Problem of the Development of Style in Later Art*. Dover.