Reducing Bias in AI-based Analysis of Visual Artworks

Zhuomin Zhang[1], Jia Li[1], *Fellow, IEEE*, David G. Stork[2], *Fellow, IEEE*,
Elizabeth Mansfield[1], John Russell[1], Catherine Adams[1],
James Z. Wang[1], *Senior Member, IEEE*
(1) The Pennsylvania State University, University Park, PA 16803 USA
(2) Independent Consultant, Portola Valley, CA 94028 USA
emails: zxz78@psu.edu, jiali@psu.edu, artanalyst@gmail.com,
ecm289@psu.edu, jer308@psu.edu, cda122@psu.edu,
jwang@psu.edu

***Abstract*—Empirical research in science and the humanities is vulnerable to bias which, by definition, implies incorrect or misleading findings. Artificial intelligence-based analysis of visual artworks is vulnerable to bias in ways specific to the domain. Works of art belong to a distinct cultural category that often prioritizes such characteristics as hand-craftsmanship, uniqueness, originality, and imaginative content; works of art are also responsive to diverse social and cultural contexts. Ascertaining which features of an artwork can be rightly ascribed to an objective "truth," without which the concept of bias is not even relevant, is itself challenging. Incorporating expert knowledge into machine learning applications can help reduce bias in final estimates. We review several sources of bias that can occur across different stages of AI-based analysis, protocols and best practices for reducing bias, and approaches to measuring these biases. This systematic investigation of various types of bias can help researchers better understand bias, become aware of practical solutions, and ultimately cultivate the prudent adoption of AI-based approaches to artwork analysis.**

Introduction

With the ever-expanding, open-source collections of digitized artworks and the rapid advancement of artificial intelligence (AI) and deep learning, particularly in the analysis of visual content, there have been early successes in AI-based analysis of artworks. Rigorous application of the techniques of computer vision, image processing, machine learning, deep neural networks, and AI, when guided by art-historical expertise and context, have provided researchers with insights and answers unavailable via traditional, non-computer-assisted analysis. Recent contributions in this field can be understood as falling into three general areas:

- Computational extraction of representative features or patterns from corpora of artworks for subsequent classification or analysis of style, composition, aesthetic qualities, and other characteristics [1], [2], [3].
- Semantic and iconographic analysis of works of art [4], [5].
- Computer generation of digital images for simulating a certain artistic style, hypothesizing the appearance of incomplete artworks, or creating new artworks for visualization and analysis [6], [7].

Although recent research has demonstrated the applicability and effectiveness of AI for several tasks in the analysis and creation of art, it is important to recognize the interplay of algorithmic processing and analysis by human experts and how this complex interplay introduces bias.

Generally, algorithmic processing is applied to digital reproductions of works of art (digital images or digital surrogates). Here, we focus on the area that has received the most attention from researchers: AI-based analysis of paintings and drawings. Bias may occur

Fig. 1. Automatic brushstroke extraction for *Wheatfield* (Arles, June 1888, oil on canvas, 54 cm × 65 cm) by Vincent van Gogh [2]. Painting image courtesy of the Van Gogh Museum, Amsterdam (Vincent van Gogh Foundation). The brushstroke map is provided by the James Z. Wang Research Group (The Pennsylvania State University).

at any step in such projects, from the problem formulation to the eventual performance evaluation. First, a well-defined research question is essential to ensuring machine analysis yields valid and meaningful results. An ill-defined problem statement can lead to biased or ambiguous findings. Next, bias can also be introduced during data collection and curation. There are many publicly accessible datasets of paintings collected for research use, including WikiArt, PeopleArt, SemArt, and WikiArt Emotions. All of these image collections contain a large number of digitized paintings with semantic labeling or annotations. As useful as these ready-made collections are, inconsistent image quality, unbalanced data distribution, and inaccurate labeling can affect the reliability of the experimental results.

The next step following data collection and curation is feature extraction—often based on handcrafted features but occasionally based on features learned from the data itself. Handcrafted feature extraction can easily introduce human bias into the process at this point. For paintings, typical features include color, edge, texture, shape, and arrangement of brushstrokes (Fig. 1) [2]. Then a trained classifier or some unsupervised method (e.g., clustering) is applied to perform various tasks, for instance, content and style representation, classification, and content recognition. Recently, the emergence of Convolutional Neural Networks (CNNs) [8] and Generative Adversarial Networks (GANs) [9] has revolutionized the task of feature extraction by learning the mapping between the input data and the ground truth labels or between different data domains directly [10]. Despite their improved performance, these methods can nevertheless suffer from biases because the true probabilistic model for the input and output is unknown [11]. To address this, the selection of representative samples for testing and choosing objective evaluation methods are also indispensable for a less biased evaluation result.

Here we use “bias” to mean, in a broad sense, systematic errors, a meaning only indirectly related to its rigorous use in mathematical statistics. In statistics literature, bias usually refers to the expectation of differences between a true parameter and its estimate. The more general sense of bias is useful for assessing the application of AI to visual arts research because the accurate analysis of bias is essential to building as faithful a model as possible. To summarize, there are several types of biases in the AI pipeline for artwork analysis. These include unreasonable problem formulation (problem formulation bias), inappropriate data curation (imaging bias, sampling bias, and labeling bias), algorithms trained with biased priors (confounding bias and design bias), and

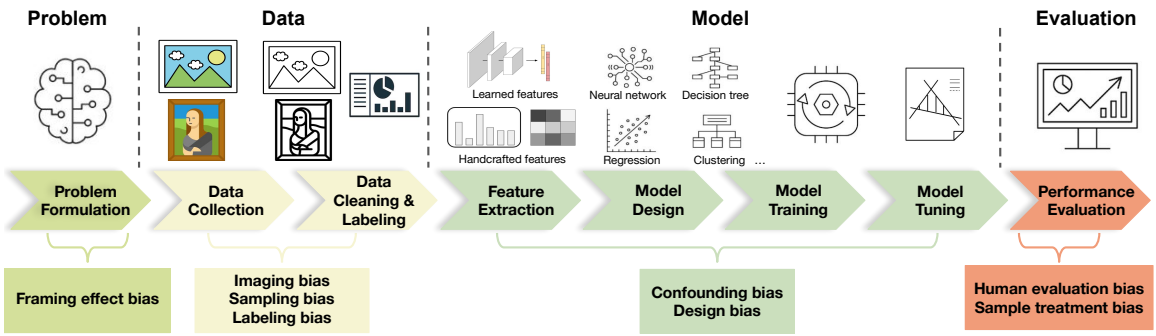


Fig. 2. Different types of bias that can occur across the stages of a pipeline for AI-based art analysis.

prejudiced evaluation methods (human evaluation bias and sample treatment bias). The terminology for the types of bias is based on previous work [12] but slightly adjusted to cater to the particular demands of art analysis. These biases could undermine the reliability of any proposed methods by generating unrepresentative features, stereotyping artists' styles, failing to capture what the artist tried to convey, and so on. Biased analysis of artworks could negatively affect research efforts to answer art-historical questions and preserve cultural heritage.

We now consider closely various types of bias across different phases of the AI-based artwork analysis pipeline from the perspective of both computer vision and art history. We begin with a survey on the sources of biases, the quantification of bias, and the methods to reduce bias, illustrating a few case studies.

Biases and Their Mitigation

The basic procedure for AI-based analysis of visual artworks consists of problem formulation, data collection, data cleaning, data annotation, algorithm design, model building, and evaluation, as illustrated in Fig. 2. Below, we describe examples of potential sources of bias at each step of the illustrated pipeline and review some methods for their mitigation. A summary of methods for reducing different kinds of bias is provided in Table I.

Problem Formulation

Bias can arise through an inappropriate formulation of a research problem, which we call *framing effect bias*. For instance, there are many ways to define the "style" of paintings since the notion of style is naturally difficult to define or quantify. A severe bias can be introduced when researchers intentionally or unintentionally define style to favor the ultimate conclusion they plan to draw.

A related situation arises when measuring pictorial similarity. For example, Zhang *et al.* evaluated paintings by John Constable, a 19th-century British artist noted for his faithful depictions of the sky, by measuring the similarity between paintings and photographs of clouds. This comparison was motivated by art-historical disagreement about the basis for the renowned accuracy of Constable's renderings of clouds, possibly due to his familiarity with Luke Howard's cloud taxonomy or strictly based on empirical observation. To assess the degree to which Constable's clouds correspond to Howard's typology, Constable's cloud paintings were compared to photographs of the same type of cloud, with photographs here serving as a reasonable ground truth for the actual appearance of a cloud. Zhang *et al.* also used the trained model to assess the relative realism of clouds by Constable and several of his contemporaries [13]. However, it is perhaps impossible to reach a consensus for judging the similarity between two paintings of clouds (Fig. 3). Conventional similarity metrics for common photographs, based on features such as color distributions, texture, layout, shape, Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HoG), cannot

TABLE I
MITIGATION METHODS FOR VARIOUS TYPES OF BIASES.

Bias Type	Mitigation Methods	Example Research Work
Framing effect bias	Multi-model learning / Boosting	Pictorial realism evaluation [13]
Imaging bias	Uniform imaging Not relying on color	Brushstroke style analysis [2]
	Virtual cleaning	Reflectance modeling [14]
	Color restoration	Color match [15]
Sampling bias	Label propagation	EKG-based artwork classification [16]
	Resampling	Dataset resampling [17]
	Data augmentation	Painting style transfer [18]
Labeling bias	Triplet generation	Triplet labels for style similarity [19]
	Few-shot learning	Contrastive learning [20]
Design bias	Feature selection	Brushstroke analysis [2]
	Using residual connection	VGG & ResNet in style transfer [21]
	Method integration through maximum-likelihood estimation	Estimates of location of illumination [22], [23], [24], [25]
Human evaluation bias	Objective quantification	Evaluation of generated captions [3]
Sample treatment bias	Fair cross-validation	Spatial cross-validation [26]

be used to compute the similarity between two cloud paintings because they do not effectively capture the underlying weather conditions depicted through the cloud paintings. For example, two cloud paintings may share similar color distributions but can represent different types of clouds.

In the above instance, the researchers attempted to reduce bias by avoiding defining similarity based on heuristics, i.e., researchers' previous experience in dealing with similar image analysis research problems. Instead, the similarity between paintings and photographs is evaluated from two aspects: whether the painted clouds can be accurately classified into corresponding cloud types and the style similarity between paintings and real photos, as shown in Fig. 4. Utilizing two different standards to realize the evaluation from different aspects can make the computation of this otherwise poorly defined similarity more reasonable and trustworthy. First, a supervised CNN model to classify cloud pictures into different types of clouds (i.e., cirrus, cirrocumulus, cirrostratus, altocumulus, altostratus, stratus, stratocumulus, nimbostratus, cumulus, and cumulonimbus) is trained using photographs of real-world clouds with labels provided by an expert meteorologist. The rationale is that if clouds depicted in paintings can be accurately classified by the model, they are likely realistic and similar to photographs. Additionally, a CNN encoder can be trained to evaluate the "style" similarity between a group of paintings and a group of real-world cloud photographs. Although the findings using these two standards are not consistent with each other, formulating this problem from different aspects helps to draw a less biased and more reliable and trustworthy conclusion.

Data

The performance of data-driven AI systems depends crucially on the scale, quality, distribution, and labeling strategies of the collected data. Here, we categorize data-related biases into three types and show example remedies in the art analysis domain to address the perfor-

Cloud Study, Hampstead, Tree at Right, 1821, by John Constable

Rome: Study of a Cloudy Sky, by Pierre Henri de Valenciennes

View of Barges on the Thames with Henley-on-Thames Beyond, 1830, by Frederick W. Watts

Dedham Water Meadows by Lionel Constable

Port of Le Havre, 1886, by Eugène Boudin

A Windy Day, 1850, by David Cox

Fig. 3. It is extremely difficult to define reliable pictorial similarity metrics among cloud study paintings. Examples by six different painters are shown. The art-historical question is: did John Constable's clouds appear more life-like than those of his contemporaries [13].

mance degradation of trained models.

Imaging bias

An overarching problem for AI-based analysis of art derives from the assumption that digital images are accurate representations of the artworks themselves. Such an assumption can lead to uncritical acceptance of images of artworks, particularly for images taken from the Web rather than from museum imaging studios. Even the most carefully created digital image of an artwork relies on properties such as lighting, cropping, staging, choice of color space, digital quantization, and so on [27]. Additionally, works of art are made of materials that change over time: they fade, change color, crack, and lose pieces; they are also altered, repaired, or cleaned by former and current owners [14]. While the analysis of color or brightness is often the most problematic, AI researchers need to be mindful of how artworks are not static objects and that digital images of artworks are the results of human and computational actions. Classic methods for reducing imaging bias either directly alter pixel color values, for example, through the use of a standard color calibration chart when photographing and brightness normalization in preprocessing, or exploit hyperspectral imaging to collect more information [28].

Color distortion is the main source of imaging bias that occurs during painting digitization. Palomero *et al.* pointed out that the "dirt" layer covering a painting, which contains oxidized varnish, dust, and faded pigments, can cause a digital image of a painting to be discolored or even appear incomplete after digitization [15]. A neural network is trained to learn the color transformation between the degraded regions of a painting (dirt-covered segments) and the regions that maintained the original appearance due to the protection of the frame (clean segments). This method successfully cleaned Fernando Amorsolo's *Malacañang by the River* digitally.

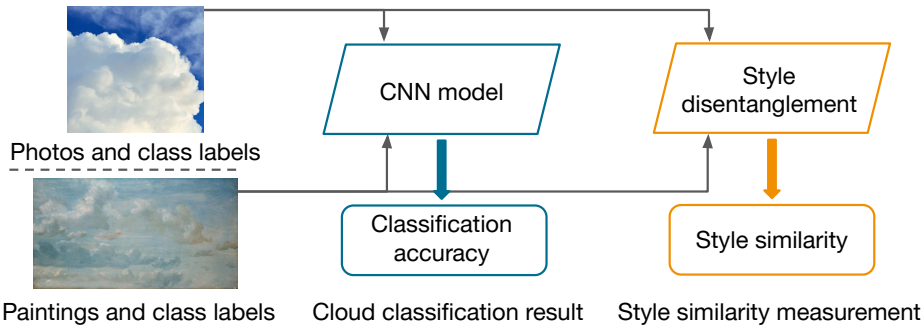


Fig. 4. An approach for evaluating pictorial realism based on classification accuracy and style comparison using a CNN-based learning framework [13].

In addition, to avoid imaging bias, one can photograph artworks under consistent conditions and imaging protocols or not rely on color in the computational analysis. For instance, Li *et al.* developed their methods for comparing Vincent van Gogh's brushstroke styles with his contemporaries based on brushstrokes automatically extracted from monochrome photos of the paintings, all digitized to a uniform density of 196.3 dots per painted inch [2]. Figure 1 shows an original painting and the brushstroke map created by the computer. Rhythmic brushstroke patterns can be clearly seen.

Virtual cleaning has been developed as a way to mitigate imaging bias related to color. Specifically, virtual cleaning is an optical simulation process to reveal the original appearance of a painting. Hyperspectral imaging techniques have enabled the capture of a set of images in contiguous narrow spectral bands so that researchers can non-destructively identify pigments, mine painting materials, and study optical changes caused by varnish discoloration and pigment fading. Aged varnish and other layers on the painting exterior are removed computationally from the digitized image. Virtual cleaning of paintings becomes possible when the distribution of materials on the surface of a painting is known. Virtual cleaning can reduce imaging biases such as inauthentic color and distorted brightness.

Trumpy *et al.* proposed a physical model for the change of diffuse reflectance to digitally remove the visual effect of aged varnish [14]. Specifically, let R_U denote the diffuse reflectance of the uncleaned painting covered by aged varnish. Here R_U is the sum of three reflections, from 1) the juncture between the air and the varnish layer, denoted by R_V^i, 2) the varnish body, denoted by R_V^b, and 3) the painting body, denoted by R_P^b. The reflection at the interface between air and the paint of the cleaned painting is denoted by R_P^i. The transmittance T of the varnish layer is estimated from the change in diffuse reflectance from the "bright" and "dark" regions of the painting before and after cleaning. Then, based on the two-flux approximation, the diffuse reflectance R_C of the cleaned painting is obtained by the sum of the reflections R_P^i and R_P^b:

$$\begin{aligned} R_C &= R_P^b + R_P^i \\ &= \frac{R_U - R_V^b}{T^2 + R_V^b(R_U - R_V^b)} + R_P^i \,. \end{aligned}$$

Although the above methods help reduce imaging bias, they may introduce new types of bias themselves if applied uncritically. Researchers thus must carefully assess each situation and choose appropriate methods accordingly.

Sampling bias

Sampling bias, also called *representation bias* [17], arises when collected data do not obey the distribution of the whole population. The deviation of the empirical data distribution from the true distribution exists for many reasons including intrinsic randomness in data and systematic issues in sampling. The former can be addressed by increasing the data size, while

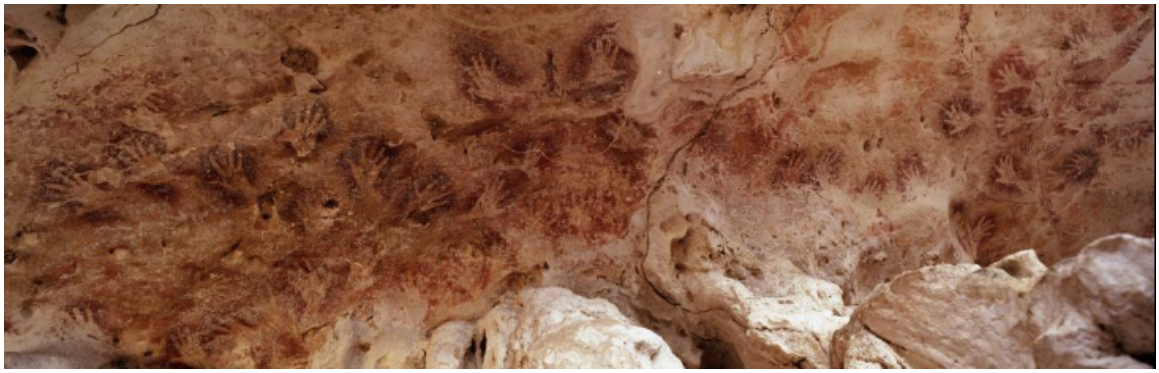


Fig. 5. 200 hand stencils on the wall of Gua Tewet, Borneo [29]. To use AI to help determine the biological sex of the artists, researchers used modern handprints as training data because it is not possible to acquire labeled handprints from people of prehistoric times.

the latter is what we usually consider as the cause of sampling bias.

One extreme case of potential sampling bias is encountered in the study of handprints on prehistoric cave artworks, dated to be over 10,000 years old (Fig. 5). The goal is to determine the biological sex of the artists [29]. Since it is impossible to collect handprints with gender labels from prehistoric humans, researchers used labeled handprints of 21st-century humans to train a model for distinguishing male versus female handprints. The underlying assumption is that the characteristics of human hands have hardly changed since the Upper Paleolithic era, a plausible assumption in light of the short span of $10,000$ years as far as evolution is concerned. On the other hand, one may argue that serious sampling bias can occur because the training data comes from a different population.

A more common situation of sampling bias arises when the numbers of instances in each class are highly unbalanced due to, for instance, the high cost of data collection, an incorrect sampling procedure, or simply the lack of annotation labels for certain classes. In the analysis of fine art paintings, we have observed imbalances in terms of painting materials (e.g., many more oil paintings than watercolors), art movements (e.g., Baroque, Impressionist), and so on.

Biases caused by unbalanced classes can be mitigated by typical methods including resampling [17], data augmentation [18], and label propagation [16]. For example, Li *et al.* proposed a pipeline to reduce sampling bias by changing the weights of samples [17]. Their strategy consists of assigning larger weights to wrongly classified instances and then optimizing the weight distribution of data points such that the classification loss is minimized. As another example, Lin *et al.* and Heitzinger and Stork exploited style transfer as a means of data augmentation to reduce sampling bias [18], [30]. In such studies authors transformed the style of photos into the style of paintings and used the transformed images as paintings.

Finally, label propagation is used to address the shortage of human annotations for collections of artworks. El Vaigh *et al.* proposed to propagate labels from annotated instances to unlabeled ones using a "Knowledge Graph" (KG) [16]. The researchers argued that label propagation and transductive learning can boost classification performance for small datasets or inadequate labels.

Let P_{labeled} be the set of labeled paintings and $P_{\text{unlabeled}}$ the set of unlabeled ones. Denote by L the set of all possible labels assigned to paintings in P_{labeled}. Denote by W all the edges between a painting and a label. For instance, an edge between a painting and the label of a certain artistic genre means this painting belongs to this genre. In addition, when based on conventional art-historical classifications, certain labels are subcategories of some other labels. Such relationships are also represented

by edges connecting two labels. Denote the set of edges between labels by R. Finally, the KG is denoted by $G = (V, E)$, where V is the set of nodes and E is the set of edges. In particular, $V = P_{\text{labeled}} \cup L$, and $E = W \cup R$.

Specifically, the researchers proposed to build an Extended KG (EKG) to construct labels for the originally unlabeled paintings. Denote the EKG by $G' = (V', E')$, where $V' = P_{\text{unlabeled}} \cup V$ and $E' = L' \cup E$. The pseudo-labels L' are obtained by a pre-trained classifier applied to those unlabeled instances. Then the graph G' is updated iteratively using a Graph Convolutional Network (GCN). The original labels stay the same while the pseudo-labels are updated when the graph is refined.

Next, we discuss another type of bias—labeling bias. As both sampling bias and labeling bias are closely related to the labels of instances, they might be confused with each other. We note that the former is concerned with an inadequate number of instances in a certain class while the latter is concerned with wrong labels given in the data, for instance due to faulty human annotation.

Labeling bias

Labeling bias usually results from subjective opinions or limited knowledge of different annotators. Factors such as social location, cultural milieu, and relevant art-historical knowledge can affect the decisions of the annotators. For example, in the WikiArt Emotions dataset, the emotion labels for each painting were given by ten annotators, which can vary somewhat. To reduce labeling bias, two approaches have been developed. In the first approach, manual labels are collected but under certain unconventional setups [19]. In the second approach, zero-shot learning [31], [20] is used to generate labels in replacement of manual labels. We next explain the two approaches in detail.

Consider the learning task of identifying feature representations that can characterize similarity between paintings based on class labels. Deep Neural Networks (DNNs) are often used to extract the features. However, due to subtleties in human perception of similarity, it is inaccurate or at best crude to assume that paintings in the same class (whether genre, art movement, or artist) appear similar while those in different classes do not. More specifically, one apparent pitfall of the assumption is that paintings in the same class are regarded as equally alike, but we often perceive a considerable amount of variation in similarity for different pairs of paintings in the same class.

To address the limitation of relying solely on class labels, Shaik *et al.* proposed to exploit additional annotations when they extracted features to capture the styles of portraits [19]. They used triplets representing style similarity provided by a professional art historian. Let (I_a, I_p, I_n) denote a triplet of images containing an anchor, a positive, and a negative image. The negative image is less similar to the anchor image than the positive one. These triplets are exploited during training to supervise the extraction of features. The objective function used in training contains a term to favor small distances between the learned features of an anchor and its positive image but large distances between those of an anchor and its negative image. Thus, paintings with similar styles are encouraged to be closer in the latent feature space of the neural network.

Some painting annotation tasks are difficult for a non-expert annotator, in which case a trained machine learning system may produce more accurate annotations. Although machine annotation is a strategy to mitigate the issue of labeling bias, researchers need to be cautioned that an automatic annotation system can introduce new biases due to its limited accuracy. The performance of the system is key to the validity of this strategy. We use the work of Conde *et al.* [20] as an example. The goal is to generate artistic attributes for each artwork in the iMet Collection Dataset [20]. Here, an attribute is a text tag that falls into five categories: country, culture, dimension (i.e., size), painting medium (e.g., watercolor, oil), and miscellaneous (i.e., all the others). For example, the description "artwork from Japan, made of paper, big size. related with woman, party, Edo" contains attributes "Japan" (country), "big size" (dimension), and "made of paper" (medium). It

is difficult for a non-expert annotator to provide such a detailed description of a painting. Conde *et al.* [20] proposed the use of contrastive learning to generate a mapping from images to text descriptions. Their system was trained using images that have been assigned with text descriptions.

Model Design

Confounding bias

In some artwork analyses, researchers try to discover causal relationships among quantities. A common bias encountered in causal discovery is called the *confounding bias*, which arises if common factors with a causal relationship with both inputs and outputs are not accounted for in the model. Suppose factor X is a common cause of Y and Z, often called a "confounder" of Y and Z. A strong correlation between Y and Z may exist due to X.

Consider confounding bias in the task of capturing the style of an artist. The "art movement" variable, denoted by X, is a confounder for the "artist," denoted by Y, and the "artwork," denoted by Z. The causal effect of artist Y on artwork Z reflects the artist's influence on the artwork, defined as the "style" of this artist. As art movement X affects the relationship between Y and Z, a reasonable approach should study the probabilistic relationship between Y and Z separately under every art movement. To accurately model the causal effect of Y on Z, every art movement in which the artist participated must be considered in order to avoid confounding bias. For example, we noticed that a proposed model, ArtGAN [32], seemingly successful at capturing and imitating artists' styles, overlooked confounders such as "art movement" and "emotions" when modeling artists' painting styles. It is thus unlikely that the "style" identified for an artist accurately represents the causal effect of the artist on the artwork.

Design bias

Design bias refers to the intrinsic deviation of a machine learning model from the optimal formulation of the prediction function. If considered under a probabilistic framework, the optimal prediction function is determined by the true joint distribution of the input and output. A variety of choices in algorithm design at various stages can cause design biases. For instance, feature extraction and selection, neural network architectures, hyperparameters of neural networks, and optimization objective functions all affect the prediction model and thus are sources of bias. Classic methods for reducing design bias include but are not limited to bagging and boosting [33], feature selection [2], and model refinement [21].

The trade-off between bias and variance to minimize average estimation error is well known in statistics. A reduced model bias comes at the cost of higher variance. Hence, a biased model is not always detrimental to accurate prediction. In fact, the widely used shrinkage techniques in statistical parameter estimation introduce bias into the model. Whether a bias is problematic depends on its severity, or, more precisely, the existence of effective control of its extent. When the dataset is small, we may have no option other than a biased model. For instance, we may rely on manually defined easy-to-explain features to carry out an analysis. Although such features reflect the subjective opinions of the researchers, their explainability is crucial for validating results in the case of small datasets.

Next, we describe a classic example of analyzing artworks in a small collection using highly explainable features. The aforementioned study on van Gogh's styles [2] shows a case of accepting bias likely to occur in feature extraction in order to maintain the validity of the analysis on a small dataset.

Motivated by the work of art historians, Li *et al.* [2] hypothesized that van Gogh's brushstrokes differed statistically from his contemporaries'. Although van Gogh's brushstrokes often strike an immediate impression on a viewer, as with many computer vision tasks, what seems apparent to the human eyes is hard to define computationally. Automatic extraction of brushstrokes poses an enormous technical challenge. The researchers developed an algorithm to extract brushstrokes and then defined eleven features to capture the geometric characteristics of individual brushstrokes, for

example, broadness homogeneity, straightness, and size, as well as characteristics of the spatial arrangement of brushstrokes, such as the number of brushstrokes with similar orientations in a neighborhood and orientation standard deviation for brushstrokes in a neighborhood. Experiments show that one set of features best distinguishes van Gogh's work from his contemporaries and another set best distinguishes his early and late periods. The distinction between van Gogh and his contemporaries lies mainly in the arrangement of brushstrokes—van Gogh's brushstrokes are more rhythmic (organized but not homogeneous throughout the canvas). Rhythmic brushstrokes are a hallmark of van Gogh's work across his early and late periods. The distinction of his work in different periods relies instead on the appearance of individual brushstrokes. These findings are easy to explain and hence can either back up the beliefs of art historians, provide new insights, or both.

Design biases tend to degrade performance when models for one task are used imprudently for different tasks. For example, Wang *et al.* noted that the residual connection, which is a crucial trait of the Residual Network (ResNet) [34] for outperforming the Very Deep Convolutional Network (VGG) [35] in some traditional computer vision tasks, is a poor design choice for image style transfer and should be omitted [21]. As another example, Liu *et al.* found that removing instance normalization in the content encoder generated worse content representations [36]. Generally speaking, models proposed for artwork analysis are task-specific and do not perform well when adopted directly for different tasks. A typical case of task shifting occurs with the change in data collection, for example, oil paintings versus watercolors, or earlier versus later periods of an artist. Transfer learning and domain adaptation approaches have been developed to alleviate such design biases. Next, we discuss a few cases which counter design biases in other ways.

Different techniques for estimating some property of an artwork—the identity of its author, the location of its central vanishing point, the location or distribution of illumination throughout a depicted tableau, and so forth—may have inherent biases. An extremely widely used technique in reducing the bias of a final overall estimate is to use *multiple different estimation methods* and combine their results in a statistically principled way. If the estimates are biased in different directions, taking the average of them typically reduces the bias. If the estimates are statistically independent, assuming that they have roughly equal variance, taking the average of them will reduce variance.

An example of such a method concerns claims about the lighting within the tableau in *Christ in the Carpenter's Studio*, a work by the Baroque master from the Lorraine, Georges de la Tour. David Hockney proposed that some Western artists, as early as 1430, secretly used optical projectors during the execution of their works and adduced de la Tour's painting as evidence [37]. The simplest computational method for estimating the location of the point-source illuminant within the tableau is cast-shadow analysis [23], [22]. The researcher simply identifies points on occluders and points on their corresponding cast shadows and draws a line between them. Although the painting is compellingly realistic, there is no reason to assume that this hand-made painting is slavishly mimetic to the scene, obeying the physics of optics. As such, we can assume that some of the cast-shadow estimates for the illuminant will be too high, and others too low.

A second method is based on the pattern of light over the floor. We assume the floor is diffusely reflecting, or Lambertian, and compute an appearance model of the brightness over the floor as a function of the (as yet unknown) location of the illuminant [23]. This appearance model has three unknown parameters, the albedo or reflectivity of the floor, the tip angle of the floor, and the two-dimensional location of the source. We estimate these parameters using gradient-descent. Such analysis shows that it is most likely that the illuminant is near the candle, surely *not* outside the frame of the painting.

A third method is a general occluding-contour algorithm [24], which was first applied in art analysis to Johannes Vermeer's *Girl with a Pearl Earring* [38]. This algorithm estimates the direction of illumination that best explains (in a sum-squared-error sense) the pattern of lightness along the outer boundary or occluding contour of an object. In *Christ in the Carpenter's Studio*, such contours included those on the faces of the figures, their arms and legs, and chests. Each such contour gives an estimate, represented as an arrow, and the best estimate of the source of illumination within the painting is the position where these lines cross most closely. There is an extremely strong agreement of the lines at the position of the candle in the tableau.

A fourth method for estimating the location of the illuminant—more specifically deciding whether the illuminant is in place of the depicted candle or instead "outside the picture"—relies upon creating a computer graphics model of the tableau [25]. During the creation of the computer graphics model of the tableau, the researchers crafted the model to match the three-dimensional *geometry* but not the effects of lighting, thereby avoiding biasing the model on the question at hand. This method avoids bias in lighting [39]. Next, the researchers adjust the position of the virtual illuminant within the tableau to candidate locations for testing the claim at hand. In the current case, the researchers place a point source *at the geometric location of the candle* and again *in place of one of the figures* (for instance Saint Joseph), as explicitly claimed. Then they visually compare the rendered virtual tableau with the artwork (for cast shadows, highlights, and so on) to see which of the two light location hypotheses is best consistent with the image in the painting. In this way, the source of illumination was found to be most consistent with the position of the candle.

Once we have estimates from multiple methods, we must integrate them to yield a final estimate in service of answering our art-historical question. A leading principled integration method is through maximum-likelihood estimation.

Evaluation

Human evaluation bias

For some learning tasks, the evaluation of a trained system requires user studies (sometimes called perceptual studies). As with the annotation process, subjectivity can be a problem during the evaluation process. The evaluator's knowledge and cultural background, among other factors, can influence their judgment, sometimes to the extent that their assessment is significantly different from the average opinion of an anticipated population. Aesthetic evaluation of visual arts is especially sensitive to individual differences, which has been observed by Kao *et al.* [40]. Bias in evaluation is sometimes caused by inaccurate labels given to test points, and hence can be categorized as labeling bias. However, since human evaluation can be different from or more diverse than labels provided for training data points, we separate human evaluation bias from labeling bias. One common idea to avoid human evaluation bias is to favor more objective and quantitative metrics for evaluation if these metrics are relevant to the art-historical question at hand. However, in some applications, objective metrics can hardly reflect users' experience, the very reason for using human evaluation despite the potential bias.

An example of mitigating this bias can be found in ArtEmis, a large-scale dataset containing paintings with annotated emotion labels and explanations for why the annotator assigned any emotion label [3]. The researchers used this dataset to train a set of neural speakers capable of generating captions to predict emotions triggered by visual stimuli, along with automatic explanations. Instead of using human annotations as the "ground truth," they trained an emotion classifier using the data. As part of the evaluation for the trained neural speakers, the deduced emotion from a generated caption is compared with this machine-generated "ground truth." The advantages of this approach are its attention to the mitigation of human bias and the availability of labeled test instances. However, an interesting question is whether the machine-predicted labels are accurate enough

for a reasonable evaluation.

Sample treatment bias

It is a common practice to split data into a training set and a testing set when estimating a machine learning model. The former is used for model estimation, while the latter is for evaluation. Data must be fully shuffled before being separated into two groups, a common practice to avoid sampling bias and evaluation bias. Otherwise, the evaluation of the model can be overly pessimistic or overly optimistic. We call the bias caused by mistakes in the usage of the sample dataset *sample treatment bias*.

Another common way to reduce sample treatment bias is to conduct cross-validation based on a random split of the data into multiple folds. When data are not independent samples, splitting them into training and testing sets for the most accurate final performance can be subtle. In an interesting work by Salazar *et al.* [26], this problem was investigated for spatially correlated data. It is found that ignoring the spatial information can cause considerable bias in evaluation, in particular overoptimistic assessment due to the spatial correlation between the test and training data. They proposed a method for splitting data to avoid such problems.

Quantifying Effects of Bias

A common way to examine the effect of bias in AI-based artwork analysis is to evaluate how the final performance changes when the bias is removed from the model, referred to as an *ablation study*. The difference in experimental results with or without the bias estimates its magnitude. Specifically, we may obtain test results with or without label propagation, certain restrictions, or changes in the network structure. We describe below an approach in a particular context to measure the effect of bias more directly than comparing the final results.

As shown in Fig. 3 of the paper by Srinivasan *et al.* [41], the style of an artist is affected by several factors, which fall into two categories: concrete variables (e.g., art material, art movements), and abstract variables (e.g., religions, emotions). Confounding biases can be expected to arise when variables with causal relationships with both artists and paintings—such as art movement, genres, and art materials—are not included in the analysis. They proposed a direct measure for the effect of art movements. By training a classifier for art movements, the authors identify features pertinent for characterizing the art movement to which a painting belongs. Denote this set of features by $\mathcal{F}$. To measure the effect of art movements on a particular artist, say artist A, A's paintings are compared with other artists' paintings in the same art movement. In the meantime, A's paintings are compared with paintings artificially generated by GAN (trained using A's work). The comparison of paintings is based only on the features from $\mathcal{F}$, those deemed relevant for characterizing art movements. Then the ratio between the average distances is used to measure the effect of art movements. The rationale is that if the art movements do not matter for how the artists created their works, the ratio is likely small since paintings from the same artist tend to be more similar than those from different artists.

Concluding Remarks

AI-based analysis of visual artworks can provide new perspectives on the study of art and answer art-historical questions that are otherwise difficult to answer. Computers excel at finding patterns in complex data and can help art historians uncover new avenues of inquiry [42]. However, given the unique, heterogeneous nature of works of art, there are many potential sources of bias in computational methods when applied to the analysis of visual artworks. In this article, we have reviewed these sources of bias and described appropriate mitigation approaches that have been attempted in the AI research community.

Having surveyed the types of biases that can arise, we want to stress that there are unsolved challenges in the AI-based analysis of visual artworks. The results obtained from a data-driven analysis must be interpreted cautiously, taking into consideration all the assumptions and limitations along the pipeline. Researchers should rigorously evaluate any model (for example, using less-populated categories to test it) before drawing conclusions. If the testing

data match poorly with the training data, a careful discussion of limitations will help avoid misleading users (e.g., art historians and the general public). This is especially important for analyzing visual artworks because we often do not have the consistency, quality, quantity, or representativeness of data that exist for other research areas. For example, when discussing AI-based analysis of paintings, many people consider the detection of counterfeits an interesting application. However, we must realize that it is impossible to collect a representative counterfeit dataset for any painter because there can be many unknown forgers possessing skills too diverse to model. A computational approach for counterfeit detection would provide insights to the extent its "reasoning" is interpretable [43].

To properly conduct AI-based analysis, researchers need to not just be mindful of the biases discussed in this article. Working closely with art historians and other relevant professionals can help surface bias across all areas, from problem formulation to evaluation, and can help devise mitigations that don't engender new problems as part of their solution. Close collaboration with experts in different fields on AI-based research projects is an important way to mitigate bias and involving art-historical expertise early on can ensure that projects avoid flawed problem formulations that render the whole analysis useless.

Acknowledgments

The work of the Penn State researchers was supported in part by the National Endowment for the Humanities (NEH) under Grant No. HAA-271801-20. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the NEH. D. G. Stork acknowledges the Getty Research Library, where some of this work was performed. J. Z. Wang acknowledges the Amazon Research Awards program for providing gifts to support his research. The authors appreciate the constructive remarks of the reviewers and the associate editor.

References

[1] C. R. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang, "Image processing for artist identification," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 37–48, 2008.

[2] J. Li, L. Yao, E. Hendriks, and J. Z. Wang, "Rhythmic brushstrokes distinguish van Gogh from his contemporaries: Findings via automated brushstroke extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1159–1176, 2012.

[3] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "ArtEmis: Affective language for visual art," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 569–11 579.

[4] G. Kell, R.-R. Griffiths, A. Bourached, and D. G. Stork, "Extracting associations and meanings of objects depicted in artworks through bi-modal deep networks," *arXiv preprint arXiv:2203.07026*, 2022.

[5] G. Duan, N. Sawant, J. Z. Wang, D. Snow, D. Ai, and Y.-W. Chen, "Analysis of Cypriot icon faces using ICA-enhanced active shape model representation," in *Proc. 19th ACM International Conference on Multimedia*, 2011, pp. 901–904.

[6] Y. Kuang, D. G. Stork, and F. Kahl, "Improved curvature-based inpainting applied to fine art: Recovering van Gogh's partially hidden brush strokes," in *Proc. SPIE 7869, Computer Vision and Image Analysis of Art II*. International Society for Optics and Photonics, 2011, p. 78690I.

[7] T. Ružić, B. Cornelis, L. Platiša, A. Pižurica, A. Dooms, W. Philips, M. Martens, M. D. Mey, and I. Daubechies, "Virtual restoration of the Ghent Altarpiece using crack detection and inpainting," in *Proc. International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2011, pp. 417–428.

[8] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, p. 2672–2680, 2014.

[10] G. Castellano and G. Vessio, "Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12 263–12 282, 2021.

[11] S. Lee, Z. J. Wang, J. Hoffman, and D. H. P. Chau, "VisCUIT: Visual auditor for bias in CNN image classifier," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 475–21 483.

[12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[13] Z. Zhang, E. C. Mansfield, J. Li, J. Russell, G. S. Young, C. Adams, and J. Z. Wang, "A machine learning paradigm for studying pictorial realism: Are Constable's clouds more real than his contemporaries?" *arXiv preprint arXiv:2202.09348*, 2022.

[14] G. Trumpy, D. Conover, L. Simonot, M. Thoury, M. Picollo, and J. K. Delaney, "Experimental study on merits of virtual cleaning of paintings with aged varnish," *Optics Express*, vol. 23, no. 26, pp. 33 836–33 848, 2015.

[15] C. M. T. Palomero and M. N. Soriano, "Digital cleaning and "dirt" layer visualization of an oil painting," *Optics Express*, vol. 19, no. 21, pp. 21 011–21 017, 2011.

[16] C. B. El Vaigh, N. Garcia, B. Renoust, C. Chu, Y. Nakashima, and H. Nagahara, "GCNBoost: Artwork classification by label propagation through a knowledge graph," in *Proc. International Conference on Multimedia Retrieval*, 2021, pp. 92–100.

[17] Y. Li and N. Vasconcelos, "Repair: Removing representation bias by dataset resampling," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9572–9581.

[18] H. Lin, M. Van Zuijlen, S. C. Pont, M. W. Wijntjes, and K. Bala, "What can style transfer and paintings do for model robustness?" in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 028–11 037.

[19] S. Shaik, B. Bucher, N. Agrafiotis, S. Phillips, K. Daniilidis, and W. Schmenner, "Learning portrait style representations," *arXiv preprint arXiv:2012.04153*, 2020.

[20] M. V. Conde and K. Turgutlu, "Clip-art: contrastive pre-training for fine-grained art classification," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3956–3960.

[21] P. Wang, Y. Li, and N. Vasconcelos, "Rethinking and improving the robustness of image style transfer," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 124–133.

[22] D. G. Stork and M. K. Johnson, "Computer vision, image analysis, and master art: Part 2," *IEEE MultiMedia*, vol. 13, no. 4, pp. 12–17, 2006.

[23] D. G. Stork, "Did Georges de la Tour use optical projections while painting Christ in the carpenter's studio?" in *Proc. SPIE 5685, Image and Video Communications and Processing*, vol. 5685. International Society for Optics and Photonics, 2005, pp. 214–219.

[24] P. Nillius and J.-O. Eklundh, "Automatic estimation of the projected light source direction," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2001, pp. I–I.

[25] D. G. Stork and M. K. Johnson, "Estimating the location of illuminants in realist master paintings computer image analysis addresses a debate in art history of the Baroque," in *Proc. 18th International Conference on Pattern Recognition*, vol. 1. IEEE, 2006, pp. 255–258.

[26] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz, "Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy," *Journal of Petroleum Science and Engineering*, vol. 209, p. 109885, 2022.

[27] R. S. Berns, *Color Science and the Visual Arts: A guide for conservators, curators, and the curious*. Getty Publications, 2016.

[28] B. Borg, M. Dunn, A. Ang, and C. Villis, "The application of state-of-the-art technologies to support artwork conservation: Literature review," *Journal of Cultural Heritage*, vol. 44, pp. 239–259, 2020.

[29] J. Z. Wang, W. Ge, D. R. Snow, P. Mitra, and C. L. Giles, "Determining the sexual identities of prehistoric cave artists using digitized handprints: A machine learning approach," in *Proc. ACM International Conference on Multimedia*, 2010, pp. 1325–1332.

[30] T. Heitzinger and D. G. Stork, "Improving semantic segmentation of fine art images using photographs rendered in a style learned from artworks," in *Computer Vision and Analysis of Art*, D. G. Stork and K. Heumiller, Eds. IS&T, 2022.

[31] P. Madhu, A. Meyer, M. Zinnen, L. Mührenberg, D. Suckow, T. Bendschus, C. Reinhardt, P. Bell, U. Verstegen, R. Kosti, A. Maier, and V. Christlein, "One-shot object detection in heterogeneous artwork datasets," in *Proc. Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2022, pp. 1–6.

[32] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "ArtGAN: Artwork synthesis with conditional categorical GANs," in *Proc. IEEE International Conference on Image Processing*. IEEE, 2017, pp. 3760–3764.

[33] O. Menis-Mastromichalakis, N. Sofou, and G. Stamou, "Deep ensemble art style recognition," in *Proc. International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

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

[36] X. Liu, S. Thermos, G. Valvano, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, "Measuring the biases and effectiveness of content-style disentanglement," *arXiv preprint arXiv:2008.12378*, 2020.

[37] D. Hockney, *Secret Knowledge: Rediscovering the lost techniques of the old masters*. Thames & Hudson London, 2001.

[38] M. K. Johnson, D. G. Stork, S. Biswas, and Y. Furuichi, "Inferring illumination direction estimated from disparate sources in paintings: An investigation into Jan Vermeer's Girl with a pearl earring," in *Proc. SPIE 6810, Computer Image Analysis in the Study of Art*. International Society for Optics and Photonics, 2008, p. 68100I.

[39] D. G. Stork and Y. Furuichi, "Image analysis of paintings by computer graphics synthesis: An inves-

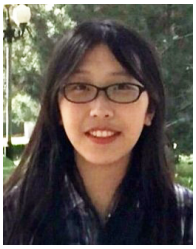
tigation of the illumination in Georges de la Tour's Christ in the carpenter's studio," in *Proc. SPIE 6810, Computer Image Analysis in the Study of Art*. SPIE, 2008, pp. 176–187.

[40] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[41] R. Srinivasan and K. Uchino, "Quantifying confounding bias in generative art: A case study," *arXiv preprint arXiv:2102.11957*, 2021.

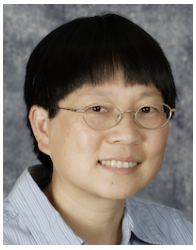
[42] E. Mansfield, Z. Zhang, J. Li, J. Russell, C. A. George S. Young, and J. Z. Wang, "Techniques of the art historical observer," *Nineteenth-Century Art Worldwide*, vol. 21, no. 1, 2022.

[43] L. Yao, J. Li, and J. Z. Wang, "Characterizing elegance of curves computationally for distinguishing Morrisseau paintings and the imitations," in *Proc. 16th IEEE International Conference on Image Processing*. IEEE, 2009, pp. 73–76.



Zhuomin Zhang received her B.S. degree in computer science from Nanjing University, China, in 2016. She is an advanced Ph.D. candidate in the College of Information Sciences and Technology at The Pennsylvania State University, University Park, PA, USA. She will join the Chronicle team at Amazon as an applied scientist in 2022.

Her research interests include computer vision, machine learning, and multimedia.



Jia Li (Fellow, IEEE) received the B.S. degree in electrical engineering (1993) from Xi'an Jiaotong University, China, and the M.Sc. degree in electrical engineering (1995), the M.Sc. degree in statistics (1998), and the Ph.D. degree in electrical engineering (1999), from Stanford University. She is a Professor of Statistics and (by courtesy) Computer Science at The Pennsylvania State University.

Her research interests include machine learning, artificial intelligence, probabilistic graph models, and image analysis. She is a Fellow of IEEE and ASA (American Statistical Association).

David G. Stork (Fellow, IEEE) received the B.S. degree in physics from the Massachusetts Institute of Technology, Cambridge, MA, USA in 1976 and the Ph.D. degree in physics from the University of Maryland, College Park, MD, USA in 1984.

He has made contributions to machine learning, pattern recognition, computer vision, artificial intelligence, computational optics, image analysis of fine art, and related fields. He is a Fellow of seven international scholarly societies and his eight books/proceedings volumes include the second edition of *Pattern Classification* and the forthcoming *Pixels & Paintings: Foundations of computer-assisted connoisseurship*.

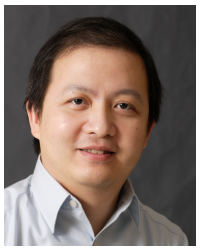
Elizabeth Mansfield received the B.A. degree in art history and the B.A. degree in linguistics from the University of California, Irvine, CA, USA, and the Ph.D. degree in art history from Harvard University, Boston, MA, USA. She is a Professor and Head of the Department of Art History at Penn State.

She is a specialist in 18th- and 19th-century European art and art historiography and her publications include *The Perfect Foil: François-André Vincent and the Revolution in French Painting* and *Too Beautiful to Picture: Zeuxis, Myth, and Mimesis*.

John Russell received the B.A. degree from the University of Vermont and the M.L.S. degree from Indiana University. He is an Assistant Professor and Digital Humanities Librarian at The Pennsylvania State University Libraries and Associate Director of the Center for Virtual/Material Studies.

Catherine Adams received the B.A. degree from The Pennsylvania State University, University Park, PA, USA and the M.L.I.S. degree from the University of Pittsburgh, PA, USA.

She worked as the assistant curator of the Visual Resources Center in the Department of Art History at The Pennsylvania State University from 2007 until 2021 when she became Digital Support Specialist in the new Center for Virtual/Material Studies.



James Z. Wang (Senior Member, IEEE) received the B.S. degree in mathematics *summa cum laude* from the University of Minnesota (1994), and the M.S. degree in mathematics (1997), the M.S. degree in computer science (1997), and the Ph.D. degree in medical information sciences (2000), all from Stanford University. He is a Distinguished Professor of the Data Sciences and Artificial Intelligence section of the College of Information Sciences and Technology at The Pennsylvania State University.

His research interests include image analysis, image modeling, image retrieval, and their applications. He was the recipient of the NSF CAREER Award (2004) and the Amazon Research Awards (2018, 2019, 2020, and 2021).