A Machine Learning Paradigm for Studying Pictorial Realism: How Accurate Are Constable's Clouds?

Zhuomin Zhang, Elizabeth C. Mansfield, Jia Li, *Fellow, IEEE,* John Russell, George S. Young, Catherine Adams, Kevin A. Bowley, and James Z. Wang, *Senior Member, IEEE*

Abstract—The British landscape painter John Constable is considered foundational for the Realist movement in 19th-century European painting. Constable's painted skies, in particular, were seen as remarkably accurate by his contemporaries, an impression shared by many viewers today. Yet, assessing the accuracy of realist paintings like Constable's is subjective or intuitive, even for professional art historians, making it difficult to say with certainty what set Constable's skies apart from those of his contemporaries. Our goal is to contribute to a more objective understanding of Constable's realism. We propose a new machine-learning-based paradigm for studying pictorial realism in an explainable way. Our framework assesses realism by measuring the similarity between clouds painted by artists noted for their skies, like Constable, and photographs of clouds. The experimental results of cloud classification show that Constable approximates more consistently than his contemporaries the formal features of actual clouds in his paintings. The study, as a novel interdisciplinary approach that combines computer vision and machine learning, meteorology, and art history, is a springboard for broader and deeper analyses of pictorial realism.

✦

Index Terms—Pictorial realism, John Constable, cloud classification, feature fusion, style disentanglement.

1 INTRODUCTION

 \prod N this paper, we propose a new machine learning paradigm for studying the European art style known as realism. The specific \blacksquare N this paper, we propose a new machine learning paradigm for case study we report here is the work of John Constable (1776- 1837) whose landscape paintings are considered foundational for the Realist movement. Constable was especially renowned for his skies. Although there is general agreement that Constable's sky paintings are persuasive in their realism, the precise basis for his realism continues to be debated. The feasibility of quantitative analysis for studying pictorial realism, as exemplified here, demonstrates that computational approaches may augment traditional approaches to art-historical research.

Fig. 1: Two *Cloud Study* oil paintings by John Constable (1822). Left: Yale Center for British Art. Right: The Frick Collection.

1.1 The Art-Historical Questions

In 1821, Constable undertook a sustained campaign of "skying," as he called his outdoor sketching of clouds. There is general arthistorical agreement that Constable's painted clouds became more life-like around this time (Fig. 1) [1], [2]. The significance of this period of concentrated effort has been debated [2], [3], [4]. Some see Constable's cloud paintings of this period as confirmation that the artist's powers of observation improved as a consequence of prolonged study, enabling him to execute more convincing clouds [5]. Yet faithful visual documentation of clouds is challenging because they are *constantly changing*. It seems reasonable to

The authors are with The Pennsylvania State University, University Park, PA 16802. Correspondence should be addressed to J. Z. Wang (jwang@psu.edu).

posit that Constable relied on certain artistic conventions or formal patterns for his paintings of these ever-shifting motifs, as painters often did. It has also been argued that the 1821 skying campaign was a belated response to the 1803 publication of Luke Howard's typology of clouds into cumulus, cirrus, stratus, etc. [6], though there is no direct evidence that Constable consulted Howard's publication [7]. Scholars remain in disagreement about the degree to which Constable relied strictly on empirical observation, on visual formulae that might escape the notice of human viewers, or on a new understanding of how to distinguish and thus represent different types of clouds [2], [3], [4]. To some extent, scholarly disagreement arises from the fact that human viewers may not perceive or may perceive only with difficulty qualities like cloud accuracy or visual conventions that have been naturalized through regular use by European artists. Our goal is to contribute to a more accurate understanding of Constable's realism via three paths of inquiry:

- 1) Do Constable's clouds correspond with the system of cloud typology introduced in 1803 by Luke Howard?
- 2) How closely do Constable's paintings emulate the appearance of actual clouds when compared to photographs of clouds?
- 3) How does the empirical accuracy of Constable's clouds compare with that of his contemporaries when judged against photographs of clouds?

1.2 Overview of Our Approach

These judgments about realism from art historians are highly subjective insofar as they record the opinion of a particular viewer at a particular moment. The perceived fidelity of a painting to the natural phenomena it represents cannot always be clearly explained, because it is guided by an immediate, intuitive response to a particular painting. This is especially true of hard-to-describe phenomena like clouds or crashing waves: for most human viewers, paintings of these subjects simply "look right" or not. To provide a more objective assessment of realism, we introduce a

machine-learning-based analysis procedure. As shown in Fig. 2, this method comprises two components: classification of painted content (cloud in this case) and evaluation of painting style. In a nutshell, we evaluate pictorial realism by assessing the similarity between paintings and photographs in terms of both the painted content and painting style, which makes our evaluation system more thorough and unbiased [8].

Fig. 2: The proposed machine learning paradigm for studying pictorial realism.

After obtaining a labeled dataset containing both photographic images of clouds and a collection of sky paintings, we first train a machine learning system using these photographs to classify cloud categories. We then apply this classifier to our painting set to predict their cloud categories. In the meantime, classification labels are created for the paintings by experts (meteorologists). The classification accuracy for the paintings is then computed and compared with the accuracy achieved for photographs. Our basic assumption is that the classification accuracy of paintings that imitate observed reality well will be close to that obtained for the photos. Further comparison can be conducted between different collections of paintings, allowing assessment of the relative fidelity of various collections to nature. One type of comparison across collections is between works by different artists. Our labeling relies on the expertise of meteorologists to categorize clouds documented in photographs and paintings according to the types defined by Howard [9]. We propose a semi-supervised learning model for cloud classification that merges classic features with edge features. The classification of clouds in Constable's paintings according to the standard typology allows for a more precise comparison with his contemporaries. By contrasting the AI system's predictions with the expert-created ground truth labels, we obtain an objective assessment of the degree to which painters are (knowingly or unknowingly) differentiating cloud types. Given the highly specialized skills and knowledge required to classify cloud types, the AI system offers an insight unattainable by the average human viewer.

Furthermore, to further explore painting styles, we examine pictorial realism from another perspective of painting style. Specifically, we first extract the encoded style features from each painter's collection by training a content-style-disentanglement model [10]. Using our newly developed evaluation metrics, we assess the pictorial realism based on these extracted style features. This allows us to compare the relative realism of various painting styles in our dataset against that of John Constable. These style features act as direct representations of the unique pictorial characteristics of each painter's collection in comparison to photographic images.

The key contributions of our work include:

• *Interdisciplinary framework:* We proposed a machine learning framework to study realism in art from an explainable and interdisciplinary perspective by leveraging computer vision techniques, meteorology expertise, and art history insights.

- *Methodology:* We developed several tools and models: a sky-ground segmentation algorithm, a new semisupervised CNN model (named SFF-CNN) for cloud-type classification, and new evaluation metrics to quantify the style differences between images. Notably, this is the first effort to harness unlabeled sky photos to enhance cloud classification.
- *Dataset:* We curated a unique dataset consisting of 363 paintings featuring skies by John Constable and six of his contemporaries. Two expert meteorologists professionally annotated each piece, making it the inaugural dataset of paintings designed for computational analysis of skies. We are sharing our sky segmentation results and detailed annotations with the broader research community.
- *Insights:* Our findings furnish the art history domain with compelling evidence: *Constable's systematic adherence to cloud typologies is pivotal for the pronounced realism in his cloud artworks.*

1.3 Related Work

We briefly introduce related work on the art-historical study of Constable's sky paintings, computerized cloud-type classification, and content-style disentanglement.

Modern art-historical scholarship on Constable's clouds began with Kurt Badt's 1950 book on the subject [7]. Prior to this, accounts of Constable's clouds were largely descriptive as opposed to analytical, attributing their realism to Constable's emotional connection with nature, his devotion to sketching outdoors, or his largely rural childhood [11]. Badt was the first to argue that Constable's proficiency with painting realistic clouds was due to his familiarity with the recent development of a typology of clouds created by British chemist Luke Howard. Howard's typology was published in 1803 and was widely disseminated during Constable's lifetime, so it was available to him. But there is no evidence that Constable possessed Howard's typology, and the artist's extant correspondence makes no direct reference to Howard [12]. More recent scholars tend to cite instead Constable's dedication to sustained periods of empirical observation of clouds [1], [5] and his familiarity with earlier paintings of naturalistic landscapes by artists like Claude Lorrain or Willem van de Velde the Younger, both of whom were well represented in English art collections during Constable's lifetime [12], [13]. In addition, a Romantic explanation for Constable's naturalism likewise persists in the scholarly literature to this day, attributing his naturalism at least in part to an emotional or spiritual impulse toward accuracy in his depictions of natural phenomena [2].

We regard the accuracy of cloud-type classification as strong evidence of Constable's familiarity with Howard's typology, so building a trustworthy cloud-type classifier is indispensable. Recently, researchers have started to adopt CNNs for cloud-type classification. Zhang *et al.* [14], [15] built a large ground-based cloud dataset, called Cirrus Cumulus Stratus Nimbus (CCSN) with cloud type labels, and a CNN model for cloud classification. Huertas *et al.* [16] proposed a feature fusion model combining CNN features and handcrafted low-level textural features to boost classification accuracy. Departing from this fusion model, our approach aims to extract more task-relevant features such as the contours of clouds to improve classification on both the photo and painting datasets.

Another problem that we address is the lack of labeled cloud photos. The emergence of semi-supervised learning can enhance classification performance by utilizing a great amount of unlabeled data during the training process. The common semi-supervised classification models can be categorized into self-learning [17], co-training [18], graph-based semi-supervised learning [19], and semi-supervised supported vector machine [20]. Following the idea of self-learning, we generate pseudo labels (detailed in Section 2.1.3) for two unlabeled sky photo datasets and then add these new data to the labeled CCSN dataset to achieve dataset expansion.

Content-style disentanglement has been extensively applied for feature decoupling, with both the content and style feature representations useful for downstream problems, such as semantic segmentation [21], [22], image retrieval [23], [24], and image style transfer [25], [26]. In image translation, most CNN-based methods aim to learn latent space representations by extracting content or style information using autoencoder variants. However, utilizing these disentangled features for similarity or discrepancy comparison among paintings from different artists–as we have done in this study–is a relatively uncharted territory.

(a) John Constable, *Study of Sky and Trees*, 1821

(b) Eugène Boudin, *Etaples, les Bords de la Canche*, 1891 Fig. 3: Sky and ground segmentation illustrated with two paintings. Left: Original paintings. Right: Homogeneous patches (represented by different colors) generated using the A3C algorithm. Regions within the thin white contours are the sky regions after regression.

2 ALGORITHMS

As we have discussed in Section 1.2, our paradigm for studying pictorial realism (Fig. 2) provides a novel perspective for comparing artworks with photographs and addresses the subjectivity of experts' opinions. Below, we elaborate on the technical components in the analysis pipeline.

2.1 Semi-Supervised Cloud-Type Classification

Our classification model consists of two main steps: clusteringbased sky segmentation and classification by a *semi-supervised feature fusion CNN* (SFF-CNN) model. The sky segmentation step reduces the impact of irrelevant parts of an image on classification.

SFF-CNN contains two streams of feature extraction, aptly called the *classic feature extractor* and *edge feature extractor*. The former generates features from low-level textures or patterns to high-level object-related characteristics, while the latter focuses on edge information. The fused features from the two encoders are utilized together for the ultimate class label prediction. We are motivated to extend a typical CNN model by incorporating edge features because (1) the contour information of cloud bases and updraft turrets is valuable for meteorologists to determine the cloud type, and (2) CNN models tend to focus on texture rather than shape for recognition [27] while paintings and photos have different texture characteristics. Our extended CNN model is trained iteratively by generating pseudo labels for unlabeled images and then refitting the model.

2.1.1 Sky Segmentation

The land, mountains, or other irrelevant regions in a painting can negatively affect cloud classification. Because only sky regions are used in the training photos, we eliminate the impact of other irrelevant parts in the paintings by excluding pixels outside the sky region from subsequent classification analysis. Specifically, a painting is segmented into two classes: sky versus non-sky (mostly land). The entire non-sky region of a painting image is replaced by black pixels and the modified image becomes the input to the CNN model, which we refer to as the *sky-selected image*.

Our sky segmentation algorithm includes two major steps: segmentation into homogeneous patches (aka, segment) and classification of each segment into sky versus non-sky. For the first step, we used the Agglomerative Connectivity Constrained Clustering (A3C) algorithm [28]. For the second step, we perform logistic regression on the features extracted from each segment to determine whether the segment is sky or non-sky. For each segment, a 10-dimensional feature vector including location and color-based features is computed. Details about the sky detection algorithm and some example results are provided in Supplementary Materials. Fig. 3 shows the clustering results of two paintings and the sky versus non-sky classification results of the segments.

2.1.2 Cloud-Type Classification

The sky-selected images are classified into different cloud types by the SFF-CNN model. Our neural network is custom-designed for cloud-type classification by incorporating pre-learned edge features into the layers of a typical CNN model as edge information is crucial in differentiating various types of clouds. The neural network consists of a bottom stream for classic feature extraction and a top stream for edge feature extraction. The classic feature extractor aims at extracting useful features from low-level textures or patterns to high-level object-related quantities, while the edge feature extractor only captures the characteristics of edges in the same input image. Both feature extractors take the three-colorchannel sky-selected images as the input.

Classic Feature Extraction: Denote the kth sky-selected image by \mathbf{I}_k . The encoder for classic feature extraction takes the three-color-channel image I_k as the input. The first two convolutional blocks both consist of two Conv-BatchNorm-Relu layers and are followed by a 2×2 pooling layer to downsample the input feature maps (400×400) . The convolutional layers in these two blocks all have stride set to 1 and the kernel size 3×3 . The next two blocks are residual blocks with two convolutional layers with stride set to 1 and 2, respectively, and the same kernel size 3×3 . Each of these blocks spatially downsamples the input feature

maps to half of their size. The third residual convolutional module follows the same structure as the first two but sets stride to 1 for both convolutional layers. Then three fully-connected layers with feature dimensions 4096, 1024, and 10, respectively, are connected to the Resconv modules. The final layer of Softmax activation produces a distribution over the ten output probability classes for each category. Lastly, the cross-entropy (CE) loss [29] is applied to train the network.

Edge Feature Extraction: Visualization results using the Grad-cam method [30] (shown in Supplementary Materials) verified our expectation that edge information is important for classifying cloud types, which motivated our strategy to fuse edge features in the CNN. We compute the edge features by a pretrained encoder named holistically-nested features for edge detection (HED) [31]. The side-output layer of each convolution module of HED generates an edge feature map at a particular receptive field size. These maps are concatenated with those generated by the CNN at corresponding layers. The two feature maps are ensured to have the same size (horizontally and vertically) such that the features at any location on one map can be combined with features at the same location on another map before convolution. In particular, we use the same setting for the HED and CNN architectures so that at every layer, their respective feature maps are generated with the same receptive field size. The augmented feature map is the input to the next convolution layer.

2.1.3 Semi-Supervised Learning

To further enhance cloud classification accuracy, we employ semisupervised learning to leverage a large set of 9,883 unlabeled cloud photos from the SkyFinder dataset [32] and FindMeASky dataset [33]. We also apply data augmentation following the schemes of FixMatch [17]. For each unlabeled image, its flipped and shifted versions, called weak augmentation images, are created. Additionally, the so-called strong augmentation images are created by another two operations, namely, CTAugment followed by Cutout [17]. We first apply the classifier trained using only the labeled images to classify the weak augmentation images. The class that has the maximum predicted posterior probability is chosen as the predicted class (also called the one-hot pseudo label). To counter the negative effect of possibly incorrect pseudo labels, the maximum predicted posterior is compared with a pre-chosen threshold. If the threshold is not exceeded, this unlabeled image and its augmented versions will not be used further. Otherwise, the pseudo label is treated as the true label for the strong augmentation images, which we refer to as high-confidence unlabeled images. Finally, another round of training is performed using both labeled and high-confidence unlabeled images. The cross-entropy between the true class and the labeled images and between the pseudo-class generated from the weak-augmented images and the predicted class posteriors using the strong-augmented images are defined as the loss to train the model.

2.2 Style Disentanglement

In addition to comparing paintings based on how well they can be classified, we propose a methodology to assess the similarity in the "style" features of pictures. In MUNIT [10], an image is decomposed into two representations: content versus style. Both the content and style features are extracted by an encoding CNN, and they can be combined as input to a decoding CNN to reconstruct the original image. Roughly speaking, the content features capture the shared characteristics between two sets of images, whereas the style features pinpoint the unique attributes of each set. The encoders and decoders for both image sets are trained together to ensure that the content features correspond to traits shared by the two sets.

In our analysis, we treat the set of paintings of every artist as domain $\mathcal A$ and the set of cloud photographs as the reference domain $\mathcal P$. This training process yields a content encoder and a style encoder for each artist. The training algorithm generates photorealistic images $I_{\mathcal{A}2\mathcal{P}}$ from images in domain $\mathcal A$ or painting-like images $I_{\vartheta 2\mathcal{A}}$ from those in domain ϑ , an operation called "crossdomain style translation." The translation is achieved by keeping the content features but adopting style features generated for an image in the other domain. These cross-domain features are fed into a decoder to reconstruct a translated picture. The training objective function used in [10] has been modified slightly in [34] by removing the learning regression loss because the authors of the latter found that better separation of content and style can be obtained and the style and input image will be more correlated. In subsequent discussions, we will refer to the style features computed via a style encoder simply as the "style" of an image.

2.2.1 Style Similarity Between Artists

First, to evaluate style similarity between artists, we consider two sets of paintings denoted by A and B. Suppose $A = \{a_i :$ $i \in \{1, 2, ..., n_A\}$ contains n_A pictures and $B = \{b_i : j \in I\}$ $\{1, 2, ..., n_B\}$ contains n_B pictures. Denote the content and style encoder trained based on style transfer from painting set A to photo set $P = \{p_k : k \in \{1, 2, ..., n_P\}\}\$ by E_C^A and E_S^A , respectively. Likewise, the encoders for B are E_C^B and E_S^B . For an image $a_i \in A$, denote its style features computed by \tilde{E}_S^A by $F_S^{a_i}$. Similarly, for any $b_j \in B$, let its style computed by E_S^B be $F_S^{b_j}$. If A and B are similar in style, we would expect $F_S^{\bar{a}_i}$ and $F_S^{\bar{b}_j}$ to be close on average. Use the normalized square of the L^2 norm of a style feature vector to indicate the signal strength: $I_S^{a_i} = ||F_S^{a_i}||^2/d$, where d is the dimension of the style feature vector. The Mean Squared Error (MSE) between $F_S^{a_i}$ and $F_S^{b_j}$ is simply $||F_S^{a_i} - F_S^{b_j}||^2/d$. For each image $a_i \in A$, we define its average distance to images in B by

$$
D_A^{a_i} = \frac{1}{n_B} \sum_{b_j \in B} \frac{\text{MSE}(F_S^{a_i}, F_S^{b_j})}{I_S^{a_i}} \,. \tag{1}
$$

Conversely, for each image $b_j \in B$, we define its average distance to images in A as $D_B^{b_j}$ likewise. Finally, define $D_A = \frac{1}{n_A} \sum_{a_i \in A} D_A^{a_i}$, $D_B = \frac{1}{n_B} \sum_{b_j \in B} D_B^{b_j}$, and

$$
D_{\text{style}}(A, B) = \frac{1}{2}(D_A + D_B). \tag{2}
$$

The distance D_{style} is taken to measure the style difference between sets A and B.

2.2.2 Style Similarity Between an Artist and Photos

Next, we propose to use the metric "Information Over Bias (IOB)" [34] to measure the difference between the paintings of an artist and real photos. For an image $a_i \in A$, where a_i is treated as a vector, let its style feature vector be $F_S^{a_i}$. $\text{IOB}(a_i, F_S^{a_i})$ is defined to quantify the amount of information in a_i which is captured by $F_S^{a_i}$. Specifically, the informativeness of $F_S^{a_i}$ is measured by the ratio between $MSE(a_i, \tilde{a_i}')$ and $MSE(a_i, \tilde{a_i})$,

where \tilde{a}'_i is a reconstructed image from an uninformative constant
substitute style yester \mathbb{I} combined with a 's content feature yester. substitute style vector 1 combined with a_i 's content feature vector, while \tilde{a}_i is generated from the informative style vector $F_S^{a_i}$
and the same content vector. Thus, we define $IOR(a, F^{a_i})$ by and the same content vector. Thus, we define $\text{IOB}(a_i, F_S^{a_i})$ by $\text{IOB}(a_i, F_S^{a_i}) = \text{MSE}(a_i, \widetilde{a_i})/\text{MSE}(a_i, \widetilde{a_i})$. With a slight abuse
of notation, we also use $\text{IOB}(A)$ to denote the average IOB values of notation, we also use $IOB(A)$ to denote the average IOB values for the images in A, i.e., $\text{IOB}(A) = \frac{1}{n_A} \sum_{i=1}^{n_A} \text{IOB}(a_i, F_S^{a_i}).$ A lower value of $IOB(A)$ indicates that the style representation of the image is less important since a substitute default style vector can result in reconstruction with a similar level of disparity from the original image. Because the style feature vectors capture the distinct characteristics of one set of images from another set, less informative style vectors reflect a higher similarity between the two set of images. To form a basis of comparison, we also compute IOB for a mixed set containing both paintings and cloud photographs. Specifically, we first compute $IOB(A)$ for a set of paintings by an artist using the style transfer process from paintings to photographs. Then we mix images from the painting set A and the photo set P to form a new set $M = \{a_i : i \in \{1, 2, ..., n_A\}, p_k : k \in \{1, 2, ..., n_P\}\}.$ Again by the style transfer process from set M to P , we can compute $IOB(M)$. Finally, the style distance between an artist and the photographs is defined as $R_{\text{style}}(A) = \text{IOB}(M)/\text{IOB}(A)$.

3 EXPERIMENTAL RESULTS

3.1 Painting and Photo Datasets

We curated a dataset of oil paintings by John Constable (1776- 1837) and six of his near-contemporaries: Pierre Henri de Valenciennes (1750-1819), David Cox (1783-1859), Frederick Richard Lee (1798-1879), Frederick W. Watts (1800-1870), Eugène Boudin (1824-1898), and Lionel Constable (1828-1887). All of these images are either high-resolution scans of existing reproductions or digital photographs of landscape paintings with "finished" clouds or pure cloud studies.

Cloud types and detailed meteorological information for each painting in the dataset were labeled by two meteorologists with expertise in cloud classification. One annotator possesses basic knowledge of the history of European landscape painting, while the other does not. Post their initial round of labeling, the two experts reached consensus on 75.5% of the labels. They both recognized that the majority of different annotations were due to borderline cases. Following a discussion between the experts, the labels used in the subsequent experiments were mostly based on the senior annotator's annotations, while the labels of 15 paintings were in accordance with the junior annotator's opinion. Finally, an open dataset containing 363 images with detailed labeled metadata was established, which will be shared (to the extent that image licensing allows) in order to facilitate further analyses of the relation between painted clouds and actual meteorological phenomena.

We used the CCSN dataset to train the cloud classification model. The CCSN dataset contains 2,543 cloud images, in which cloud photographs were labeled into 10 cloud categories, thus we formulated cloud-type classification as a 10-class problem. For semi-supervised learning, we leveraged the SkyFinder [32] and FindMeASky [33] datasets, which came with the sky segmentation masks but no cloud-type labels. After eliminating duplicate images, our unlabeled dataset comprised 9,883 photos.

3.2 Cloud Classification on the Paintings

To evaluate our sky segmentation algorithm, we manually labeled sky regions for all 363 paintings, which serve as the ground truth. We then computed pixel accuracy, mean accuracy, and mean IoU as evaluation metrics, which were 0.9804, 0.9613, and 0.9427, respectively. Such accuracy levels are regarded as high.

Applying the trained SFF-CNN to the test photo images (20% of the CCSN dataset), we obtained a classification precision of 97.2% and recall of 96.9%. Detailed results on the test photos are provided in Section 3.5. Then, we re-trained the classification model on the entire CCSN dataset, which was then applied to the paintings. Because the painting dataset was small and the prevalence of different cloud types was highly unbalanced, to compute classification accuracy for the paintings, we only discriminated at the granularity of five common cloud types: cumuliform (cumulus), cumulonimbiform (cumulonimbus), cirriform (cirrus), stratiform (stratus, cirrostratus, altostratus, and nimbostratus), and stratocumuliform (cirrocumulus, altocumulus, and stratocumulus) [35]. The classification accuracy of each painter using the SFF-CNN model with or without feature fusion is shown in Fig. 4. For the accuracy achieved with feature fusion, the confidence interval for the accuracy at the significance level of 0.05 is shown. Except for Cox, all the other artists had a confidence interval of accuracy well above 60% (higher than the percentage of the most dominant cloud type), indicating that the clouds they painted correspond with Luke Howard's system of cloud categorization to a great extent. Moreover, clouds painted by Constable were the easiest to classify (highest accuracy) with a classification accuracy of 0.8452. Additionally, in Fig. 5, we show the classification confusion matrices for each artist's paintings. Constable's clouds achieved the highest classification accuracy in the cumuliform.

Fig. 4: The comparison in terms of classification accuracy of all seven painters using the SFF-CNN model with or without feature fusion. The error bars denote the confidence interval for the classification accuracy at the significance level of 0.05 for each painter.

To compare Constable with each of the other artists, we conducted hypothesis testing with the alternative hypothesis: Constable's paintings can be more accurately classified than those of other artists. We assigned identification numbers with Constable represented by 1 and the other artists labeled as 2, 3, ..., 7. We modeled the classification decision on a painting of the ith artist by a Bernoulli random variable with 1 indicating the correct classification and 0 otherwise. Let p_i be the probability of correct classification. Thus, the distribution for the number of correctly classified paintings of artist i is a Binomial distribution. The null hypothesis we formulated is $p_1 \leq p_i$, $i \neq 1$. We used the one-tail

Fig. 5: The confusion matrices represent the classification results of all seven painters using the semi-supervised feature fusion model. The vertical axis represents the ground truth, while the horizontal axis represents the predicted labels. The abbreviations Cu, Cb, Cs, St, and Sc stand for cumuliform, cumulonimbiform, cirriform, stratiform, and stratocumuliform, respectively.

Z-test [36] with continuity correction. The p-values obtained for Valenciennes, Lee, Watts, Lionel, Boudin, and Cox were 0.332, 0.189, 0.128, 0.147, 0.024, and 0.189, respectively. At the significance level of 0.1, Constable's paintings were more accurately classified than Boudin's works, but not others. We conducted the same hypothesis testing to examine whether the inclusion of edge features could significantly improve the classification accuracy for any artist. The lowest p-value was 0.122, obtained for Constable, while the other *p*-values exceeded 0.25 . This result indicates that the edge features improved classification most significantly for Constable.

From the results, it is evident that Constable's clouds correspond well with the system of cloud typology devised by Luke Howard. The 5% confidence interval for the classification accuracy of Constable's paintings was [0.768, 0.923]. The average classification accuracy was highest for Constable's paintings. Are Constable's clouds more reminiscent of photographs of real-world clouds than those of his contemporaries? The answer is mixed. At the significance level of 0.1, as indicated by the aforementioned p-values, Constable's clouds were more accurately classified than Boudin's, but not more than those by Valenciennes, Lee, Watts, Lionel Constable, and Cox. A potential explanation for the insignificant difference between Constable and these artists could be the limited number of paintings each of them had in the dataset.

We posit that Constable's technique, which involves strong contour lines rendered with a relatively continuous brushstroke, contributes to the realism of his clouds. In contrast, some artists, such as Boudin, tended to use dots and dashes in lieu of the clear-edged and smooth contours that define cloud shapes. Our computer model–trained on photographs–found Constable's cloud representations easier to classify and thus to recognize by viewers. Attention to precisely the morphological differences that Luke Howard highlighted when crafting his cloud typology in 1803 endowed Constable's clouds with a sufficiently striking degree of realism to set him apart from other landscape painters, at least in the eyes of his contemporaries–and in the eyes of our computer models. While our findings cannot confirm *definitively* that Constable was acquainted with Howard's cloud classification, they do confirm that systematic categorization is key for the visual impact of his realism.

3.3 Style Similarity Analysis

To train the style encoder for each artist, we used the MUNIT model [10] as the network backbone. We excluded Learning Regression loss during training as suggested in [34] for better disentanglement of content and style features. All the paintings of an artist formed set A, and a subset of cloud photographs formed set P. We selected 300 cloud photographs and ensured that the number of images in each cloud category was the same. For the paintings, instead of the original images, we used their skyselected images. After obtaining the style encoders, we computed D_{style} and R_{style} .

3.3.1 Style Distance Between Artists' Clouds and Cloud Photos

We computed R_{style} (defined earlier) for each of the seven painters. To assess variation in R_{style} caused by randomness in the input images, for each painter, we randomly sampled five paintings to form a set and computed R_{style} for this set. The calculation was repeated for multiple random samples of five paintings. As our collection only contained nine paintings by Cox, there were a maximum of 126 different combinations of five paintings by Cox. We thus randomly sampled subsets of five paintings 126 times for every artist. Table 1 shows the average values of R_{style} for each artist as well as the standard deviation.

To assess whether the distance metrics vary significantly among artists, we conducted hypothesis testing with the alternative hypothesis: these distances are significantly different between the artists. Denote the set of paintings from each of the seven artists by C_i , with $i = 1, 2, ..., 7$, and the sampled subsets by C_i^n , where $n = 1, 2, ..., 126$. Let the set $R_{\text{style}}(C_i) = \{R_{\text{style}}(C_i^n)$: $n \in \{1, 2, ..., 126\}$. Assume that the distribution of $R_{\text{style}}(C_i^n)$ for each set C_i follow a Gaussian distribution $N(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 indicate the mean and variance, respectively. Our null hypothesis is: $\mu_1 = \mu_2, \ldots, = \mu_7$. We use an F-test for a one-way analysis of variance. With an F-statistic of 21.15 and a p-value below $2e - 16$, the null hypothesis (the sets have the same mean value) is rejected at the significance level 0.05. Then, we conducted another hypothesis test using the T -test to test if the paintings of Constable exhibit a style more akin to photographs compared with other artists. Let μ_1 denote the mean value of Constable's painting set. We conducted six hypothesis tests with the null hypothesis: $\mu_1 \geq \mu_i$ for $i = 2, 3, ..., 7$. Table 1 shows both the T -statistics and the corresponding p -values. At a significance threshold of 0.1, John Constable's painting style appears more similar to photographs than that of Boudin, Lee, and Cox. However, we cannot reject the null hypothesis that his painting style is less photo-like than that of Valenciennes, Lionel Constable, and Watt. In addition, we conducted the same T-test to determine whether, on average, R_{style} of Valenciennes surpassed that of the other artists. All the p-values fell below 0.1. This result suggests that Valenciennes' painting style is the most reminiscent of actual photos when compared with the other six painters, at a significance level of 0.1. Furthermore, the Pearson correlation coefficient between classification accuracy and style similarity is -0.782, with a p-value of 0.039. This strong negative correlation

between the measurement of stylistic difference (paintings versus photos) and the accuracy of cloud classification aligns well with our heuristic understanding–paintings similar to photos tend to be classified more accurately into cloud types.

TABLE 1: R_{style} of the painting sets of each painter and T statistics of T-test about the difference of R_{style} between John Constable and other artists.

Artist	R_{style} (mean \pm std)	T -statistic	p -value
Valenciennes	1.163 ± 0.132	1.590	0.944
Lionel Constable	1.188 ± 0.141	0.165	0.566
John Constable	1.191 ± 0.147		
Watts	1.210 ± 0.146	-1.029	0.152
Boudin	1.254 ± 0.143	-3.448	3.310e-04
Lee	1.298 ± 0.151	-5.699	1.689e-08
Cox	1.319 ± 0.156	-6.703	6.775e-11

TABLE 2: Style distance D_{style} between painting set of John Constable with himself or others and T statistics of T -test about the difference of D_{style} .

3.3.2 Style Similarity Between Paintings by Constable and His Contemporaries

Next, we used Eq. (2) to compute the style similarity between pairs of painters. The results are shown in Table 2. Again, we conducted hypothesis testing to verify whether these style distances were significantly different. We used C_1 to denote the set of paintings by John Constable, and C_i for those by another artist i. Similar to the approach in the previous subsection, we computed D_{style} between randomly sampled subsets of paintings by two artists. The same subsets used to generate R_{style} were used here. For the pair of sets C_1 and C_i , we obtained 126 values of D_{style} : $D_{\text{style}}(C_1, C_i) = \{D_{\text{style}}(C_1^n, C_i^n), n \in \{1, 2, ..., 126\}\}.$ To establish a baseline, we also computed D_{style} for subsets of paintings within John Constable's collection. Specifically, in addition to the 126 subsets C_1^n that were already created, another 126 random subsets were sampled from C_1 , each containing five paintings. Denote these new subsets by $C_{1,2nd}^n$, $n \in \{1, 2, ..., 126\}$. Then, $D_{\text{style}}(C_1, C_1) = \{D_{\text{style}}(C_1^n, C_{1,2nd}^n), n \in \{1, 2, ..., 126\}\}.$ If Constable's style significantly diverges from that of other artists in terms of D_{style} , we would expect the values in $D_{\text{style}}(C_1, C_i)$ for $i \neq 1$ to surpass, at least on average, those in $D_{\text{style}}(C_1, C_1)$.

Denote the mean of $D_{\text{style}}(C_1, C_i)$ by μ'_i . In the first test, the null hypothesis is: $\mu'_1 = \mu'_2, ..., = \mu'_7$. Similarly, we used the F -test for one-way analysis of variance. The F -statistic obtained was 12.69 with a p-value of $9.21e - 14$, suggesting a significant difference in the style features among these paired artists.

The style distances between other artists and John Constable are provided in Table 2. We also conducted a T-test between two data sets $D_{\text{style}}(C_1, C_1)$ and $D_{\text{style}}(C_1, C_i)$, where $i \in \{2,...11\}$ to test if artist i's painting style is similar to John Constable's. The

null hypothesis is: $\mu'_1 \ge \mu'_i$ for $i \ne 1$. We tested at the confidence level of 0.95. The T -statistic and the corresponding p -value for the 6 tests are listed in Table 2, and we can observe that p -values are all below 0.05 except for Lionel Constable. We can therefore claim that Lionel Constable's paintings are the most stylistically similar to John Constable's of the group.

3.4 Insights for Art History

The key art-historical findings are: (1) John Constable's clouds can be more accurately classified than those of his contemporaries, which sustains the possibility that Constable possessed some knowledge of Luke Howard's classification of clouds but does not serve as definitive proof. (2) Fusing edge features boosts the classification performance of Constable's clouds more than it does for other artists. This underscores the significance of the pronounced structure in Constable's clouds as a contributing factor to their realistic portrayal. (3) John Constable's paintings are not the most realistic among the artists evaluated if realism is defined by relative approximation in appearance to a photograph. Valenciennes, according to our experiments, created clouds that bear the closest resemblance to photographs. (4) In terms of painting style, Lionel Constable aligns most closely with John Constable. This is consistent with his known practice of emulating his father's style.

3.5 Classification Results on Cloud Photos

We randomly selected 20% of the images from the CCSN dataset for testing. The other 80% of the labeled images from the CCSN dataset and all the unlabeled images were used together during the self-learning process. In the training process, only parameters in the encoder for classic feature extraction were learned by backpropagation, while the parameters of the edge feature encoder were fixed. We chose Adam as the optimizer with a learning rate of 0.0001 and batch size of 16, which provided the highest accuracy. We compared the classification results obtained by our model with two advanced methods, CloudNet [14] and ensemble-learningbased classification [16]. Our SFF-CNN model achieved the best performance with a precision of 0.972 and a recall of 0.969. The confusion matrix is shown in Fig. 6. In contrast, CloudNet (Ensemble learning) achieved a precision of 0.891 (0.953) and a recall of 0.868 (0.902). We also conducted the ablation study on the SFF-CNN model with results shown in Table 3. The improvement of the classification accuracy of SFF-CNN can be attributed to sky selection, the usage of unlabeled data, and edge feature fusion.

Fig. 6: Confusion matrix of the test results on the CCSN dataset using our SFF-CNN model.

4 CONCLUDING REMARKS

Moving beyond investigating this artistic movement solely through traditional methods of art history or via computer-aided stylometric analysis, we engage with meteorology both as a means of gaining ground truth and as a historical discipline that may have influenced visual arts.

Following the assumption that the more realistic the cloud painting is, the easier it is for the AI to determine its cloud type, we developed a new, specialized computer-based cloud-type classification method to determine if Constable's clouds or those of his contemporaries can be correctly categorized into different cloud types. Additionally, by content-style disentanglement, we defined two metrics to evaluate the style similarity between paintings and photos as well as the similarity among artists.

Further avenues for art-historical inquiry are indicated by our research. The stylistic similarity between Valenciennes and Constable invites a reconsideration of their relationship. Our experiments suggest that even artists closely associated with naturalism like Boudin were working in a less photographic mode than like-minded predecessors who died just before photography was invented. This raises the interesting possibility that a kind of photographic realism was highly prized around 1800, but was soon seen as less realistic when applied to painting once photographs were more or less ubiquitous after the 1850s. These possibilities can be investigated further using the presented style similarity analysis.

ACKNOWLEDGMENTS

This research was supported by the National Endowment for the Humanities (NEH) under Grant Nos. HAA-271801-20 and HAA-287938-22. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which was supported by the National Science Foundation (NSF) Grant No. ACI-1548562. J. Z. Wang has been supported by generous gifts from the Amazon Research Awards program and the NSF under Grant No. CNS-2234195. Xinye Zheng and Kevin R. Victor contributed in the early stage of the project. The authors would like to thank the Yale Center for British Art for hosting a visit of the team and for valuable discussions. They are grateful to the reviewers and editors for many constructive comments.

REFERENCES

- [1] J. E. Thornes and J. Constable, *John Constable's Skies: A fusion of art and science*. A&C Black, 1999.
- [2] A. Lyles, "This glorious pageantry of heaven," in *Constable's Skies*, F. Bancroft, Ed. New York: Salander-O'Reilly Galleries, 2004, pp. 29–54.
- [3] S. Cove, "Very great difficulty in composition and execution: The materials and techniques of Constable's sky and cloud studies of the 1820s," in *Constable's Skies*, F. Bancroft, Ed. New York: Salander-O'Reilly Galleries, 2004.
- [4] ——, "Clouds over Europe," in *Constable's Clouds: Paintings and Cloud Studies by John Constable*, J. Gage, Ed. National Galleries of Scotland, 2000.
- [5] G. Reynolds, "Constable's skies," in *Constable's Skies*, F. Bancroft, Ed. New York: Salander-O'Reilly Galleries, 2004, pp. 19–27.
- [6] L. Howard, "On the modifications of clouds, and on the principles of their production, suspension, and destruction: Being the substance of an essay read before the Askesian Society in the session 1802–3," *The Philosophical Magazine*, pp. 5–11, 1803.
- [7] K. Badt, *John Constable's Clouds*. Taylor & Francis, 1950.
- [8] Z. Zhang, J. Li, D. G. Stork, E. Mansfield, J. Russell, C. Adams, and J. Z. Wang, "Reducing bias in ai-based analysis of visual artworks," *IEEE BITS the Information Theory Magazine*, 2022.
- [9] L. Howard, *Essay on the Modifications of Clouds*. John Churchill & Sons, 1803.
- [10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 172–189.
- [11] A. Chamberlain, *John Constable*. George Bell & Sons, 1903.
- [12] L. Hawes, "Constable's sky sketches," *Journal of the Warburg and Courtauld Institutes*, pp. 344–365, 1969.
- [13] A. C. Esmeijer, "Cloudscapes in theory and practice," *Simiolus: Netherlands Quarterly for the History of Art*, vol. 9, no. 3, pp. 123–148, 1977.
- [14] J. Zhang, P. Liu, F. Zhang, and Q. Song, "CloudNet: Ground-based cloud classification with deep convolutional neural network," *Geophysical Research Letters*, vol. 45, no. 16, pp. 8665–8672, 2018.
- [15] J. Zhang, P. Liu, F. Zhang, H. Iwabuchi, A. A. d. H. e Ayres, V. H. C. De Albuquerque *et al.*, "Ensemble meteorological cloud classification meets internet of dependable and controllable things," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3323–3330, 2020.
- [16] J. Huertas-Tato, A. Martín, and D. Camacho, "Cloud type identification using data fusion and ensemble learning," in *Proc. Int. Conf. Intelligent Data Engineering and Automated Learning*. Springer, 2020, pp. 137– 147.
- [17] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semisupervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.
- [18] J. Fan, B. Gao, H. Jin, and L. Jiang, "Ucc: Uncertainty guided crosshead co-training for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 9947–9956.
- [19] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-based semi-supervised learning: A comprehensive review," *IEEE Trans. Neural Networks and Learning Systems*, 2022.
- [20] Z. Wan, Y. Dong, Z. Yu, H. Lv, and Z. Lv, "Semi-supervised support vector machine for digital twins based brain image fusion," *Frontiers in Neuroscience*, p. 802, 2021.
- [21] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101535, 2019.
- [22] X. Liu, S. Thermos, A. O'Neil, and S. A. Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 307–317.
- [23] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, and Y.-Z. Song, "Stylemeup: Towards style-agnostic sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 8504–8513.
- [24] S. Lee, D. Kim, and B. Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 802–812.
- [25] G. Kwon and J. C. Ye, "Diagonal attention and style-based gan for content-style disentanglement in image generation and translation," *arXiv preprint arXiv:2103.16146*, 2021.
- [26] D. Kotovenko, A. Sanakoyeu, S. Lang, and B. Ommer, "Content and style disentanglement for artistic style transfer," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2019, pp. 4422–4431.
- [27] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.
- [28] J. Li, "Agglomerative connectivity constrained clustering for image segmentation," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 1, pp. 84–99, 2011.
- [29] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via

gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 618–626.

- [31] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1395–1403.
- [32] R. P. Mihail, S. Workman, Z. Bessinger, and N. Jacobs, "Sky segmentation in the wild: An empirical study," in *2016 IEEE Winter Conf. Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–6.
- [33] S. Rawat, S. Gairola, R. Shah, and P. Narayanan, "Find me a sky: A data-driven method for color-consistent sky search and replacement," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2018, pp. 216–228.
- [34] X. Liu, S. Thermos, G. Valvano, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, "Metrics for exposing the biases of content-style disentanglement," *arXiv preprint arXiv:2008.12378*, vol. 7, 2020.
- [35] E. Barrett and C. Grant, "The identification of cloud types in LANDSAT MSS images," *ERTS Follow-on Programme Study Report, U.K. Department of Industry*, 1976. [Online]. Available: https://ntrs.nasa.gov/citations/19760014556
- [36] S. Suissa and J. J. Shuster, "Are uniformly most powerful unbiased tests really best?" *The American Statistician*, vol. 38, no. 3, pp. 204–206, 1984.

John Russell is Associate Professor and Digital Humanities Librarian at The Pennsylvania State University Libraries and Associate Director of the Center for Virtual/Material Studies. He holds a B.A. from the University of Vermont and an MLS from Indiana University.

George Young is a Professor of Meteorology at The Pennsylvania State University. His research interests are observational and statistical meteorology on scales ranging from individual clouds to large-scale weather systems. His contributions include participation in the planning, execution, and data analysis of numerous observational programs. His computational research involves the application of statistics, optimization, and machine learning to weather forecasting, satellite cloud image typing, and a range of

interdisciplinary problems. He has been a member of the Penn State faculty since 1986. He received an MS degree in Meteorology (1982) from Florida State University and a PhD degree in Atmospheric Science (1986) from Colorado State University. He is a Fellow of the American Meteorological Society.

Zhuomin Zhang is an Applied Scientist at Amazon in Seattle. She received her Ph.D. degree in Information Sciences and Technology from The Pennsylvania State University in 2022 under the joint supervision of Professors James Z. Wang and Jia Li. She received her bachelor's degree in Computer Science from Nanjing University in 2016. Her primary research interests include computer vision, machine learning, and multimedia.

Elizabeth Mansfield is Professor of Art History at Penn State. A specialist in 18th- and 19th-century European art and art historiography, her publications include the books The Perfect Foil: François-André Vincent and the Revolution in French Painting and Too Beautiful to Picture: Zeuxis, Myth, and Mimesis. The latter was awarded the Charles Rufus Morey book prize by the College Art Association in 2008. Past positions include academic appointments at New York University and at Sewanee: The

University of the South. Prior to joining the faculty at Penn State in 2018, she served as Senior Program Officer at the Getty Foundation and Vice President for Scholarly Programs at the National Humanities Center. Her current book project is a study of Realism.

Jia Li (Fellow, IEEE) 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 worked as a Program Director at the National Science Foundation from 2011 to 2013, a Visiting Scientist at Google Labs in Pittsburgh from 2007 to 2008, a researcher at the Xerox Palo Alto Research Center from 1999 to 2000, and a Research Associate in the Com-

puter Science Department at Stanford University in 1999. She received 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 was Editor-in-Chief for Statistical Analysis and Data Mining: The ASA Data Science Journal from 2018 to 2020. She is a Fellow of the Institute of Electrical and Electronics Engineers and a Fellow of the American Statistical Association.

Catherine Adams worked as the assistant curator of the Visual Resources Centre 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. She has experience in metadata and digital image creation and curation. She has managed the Art History Department Visual Resource Collection and the Palmer Museum of Art's online collections in CONTENTdm. She aided in the creation of Arts

and Architecture Resource Collaborative (AARC) and in several Omeka and Omeka S sites. She holds a B.A. from The Pennsylvania State University and an MLIS from the University of Pittsburgh.

Kevin A. Bowley received a B.S. degree in meteorology from Penn State University (2009) and a M.Sc. and Ph.D. degree in atmospheric and oceanic sciences from McGill University (2011 and 2017, respectively). He joined the faculty in the Department of Meteorology and Atmospheric Science at Penn State University in 2017 as a Lecturer, became an Assistant Teaching Professor in 2018, and is currently an Associate Teaching Professor (2023). His research and teaching interests center on understanding

topics in synoptic meteorology, with a specific focus on Rossby wave breaking processes in the present and future climate.

James Z. Wang (Senior Member, IEEE) 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. He received the bachelor's degree in mathematics *summa cum laude* from the University of Minnesota (1994), and the MS degree in mathematics (1997), the MS degree in computer science (1997), and the PhD degree in medical information sciences (2000), all from Stanford University. His research

interests include image analysis, affective computing, image modeling, image retrieval, and their applications. He was a visiting professor at the Robotics Institute at Carnegie Mellon University (2007-2008), a lead special section guest editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (2008), a program manager at the Office of the Director of the National Science Foundation (2011-2012), and a special issue guest editor of the IEEE BITS – The Information Theory Magazine (2022). He is also affiliated with the Department of Communication and Media, School of Social Sciences and Humanities, Loughborough University, UK (2023-2024). He was a recipient of a National Science Foundation Career Award (2004) and Amazon Research Awards (2018- 2022).

Supplementary Materials for "A Machine Learning Paradigm for Studying Pictorial Realism: How Accurate Are Constable's Clouds?"

✦

1 ART HISTORICAL BACKGROUND

1.1 Artists' Use of Pictorial Conventions

As noted in Section 1.1 of our paper, artists may rely on visual codes, conventions, or symbols to convey information to a viewer. Viewers accustomed to the visual codes or visual symbols of a particular culture may not even be aware of the use of such conventions. An example is the tendency for children raised in some cultures, including in the U.S., to represent the sun as a circle with a smiley face and several lines emanating from the circle outward. Although this is not a life-like representation of the sun, it is immediately recognizable as the sun by most American viewers. Visual codes can be much subtler, of course. This phenomenon is readily evident in depictions of ocean waves, which are just as recognizable when they are represented through the use of an artistic convention familiar to the viewer as they are when they are portrayed in a highly naturalistic way (Fig. 1). So one art historical explanation for the perceived truthfulness of Constable's clouds is the fact that viewers who are familiar with the tradition of European landscape paintings are accustomed to seeing clouds depicted in this way and are also accustomed to attributing to such paintings a quality of life-likeness.

1.2 How Luke Howard's Essay on the Modification of Clouds Might Have Influenced John Constable?

This argument hypothesizes that Constable came to understand, by way of Luke Howard's account [1], the nomenclature, distinct physical structures of different types of clouds, and the atmospheric conditions that generate different types of clouds, and that this knowledge enabled Constable to paint clouds more persuasively. The 1803 publication of Howard's "On the Modifications of Cloud" included verbal descriptions and visual illustrations of different cloud types (Fig. 2). Howard's nomenclature continues to be used today.

2 DATA

2.1 Painting Dataset

The key factors we used to select proper artistic works for comparison are as follows:

We should maintain a dataset that is consistent in terms of medium. Because many of Constable's most renowned

depictions of clouds were painted with oil rather than watercolor, we should find comparative works that are also oil paintings.

- It can be hard to know for certain that a cloud study was entirely executed outdoors or touched up in the studio, so we should use artists who worked out of doors as well as in the studio.
- We should use artists for whom clouds were of enduring interest. By focusing on artists whose oeuvres include many depictions of clouds, we may be able to collect a large enough dataset.

All of the artists in our dataset, worked in oil and all had a sustained interest in painting skies/clouds. For instance, Lionel, son of John Constable, emulated his father's technique; French artist Eugène Boudin was known as "king of skies" and encouraged a number of artists like Gustave Courbet and Claude Monet to paint clouds *en plein air* (*i.e.*, in the open air); Pierre-Henri de Valenciennes trained younger artists to paint out-of-doors and to practice making cloud studies. Other painters in our dataset were similarly attentive to the depiction of cloudy skies.

Fig. 3 shows the painting distribution in our dataset in terms of painters and cloud types. As can be seen, there are more paintings by Boudin and John Constable and more depictions of cumulus clouds in our dataset.

We illustrate some representative paintings of each artist in Fig. 4 and Fig. 5 to provide a general impression of these artists' landscape paintings.

2.2 Photo Dataset

The CCSN dataset [2] contains 2,543 cloud images in total. According to the World Meteorological Organization's genera-based classification recommendation, all the collected images are divided into 11 different categories as shown in Table. 1. Representative sample images from each category are shown in Fig. 6. All images are fixed resolution 400×400 pixels in the JPEG format.

To achieve semi-supervised learning, we leverage the SkyFinder [3] and FindMeASky [4] datasets to boost the classification performance. The SkyFinder dataset contains over 90,000 outdoor sky photos in different weather situations with associated detailed weather data and annotated sky pixels. However, not all photos were taken in a cloudy situation and there are plenty of

Fig. 1: Waves in art: From engraved maps and woodblock prints to contemporary photography. (a) Detail from Abraham Ortelius, *Theatrum Orbis Terrarum*, The Netherlands, hand-colored engraving, 1570. (b) Katsushika Hokusai, *The Great Wave off Kanagawa*, Japan, woodblock print, 1830-33. (c) Detail from Luis Ramos, *David Mitchell Riding a Wave*, Puerto Rico, photograph, 2015.

Fig. 2: Luke Howard's *Essay on the Modifications of Clouds*, original 1803 edition. (a) Title page. (b) First page. (c) Plate VII.

repetitive views of the exact same cloud, so we only used images labeled as "cloudy," and we eliminated images taken from the same camera and on the same day to avoid including multiple images of the same cloud. In addition, the FindMeASky dataset consists of 6,679 outdoor sky images with corresponding binary masks filtered from the ADE20K Dataset [5] where the sky region occupies over 40% of the area of the whole image. Therefore, our unlabeled dataset has 9,883 images in total.

3 THE SKY SEGMENTATION METHOD

We refer to the idea developed by J. Li [6] as the basis of our segmentation algorithm. The proposed image segmentation algorithm by Li [6] is called agglomerative connectivity constrained clustering (A3C) which combines the top-down k-means clustering and a bottom-up agglomerative connectivity constrained merge method to achieve image segmentation. In our case, we first obtain the segments through the A3C algorithm and then apply a logistic regression on the location and color features extracted from each segment to achieve a two-class sky-land segmentation.

K-means Clustering: First, we apply multi-depth k-means clustering on the LUV color space of each image to get small segmented patches homogeneous in color. Suppose K clusters are generated after k-means clustering, then a graph G recording the connectivity between clusters is built using these K clusters C_k , $k = 1, 2, ..., K$ as nodes. If there exists one pixel from C_i that is 8-connected with a pixel in C_j , we consider C_i and C_j adjacent. If C_i and C_j are adjacent, edge (C_i, C_j) exists in G, represented as $(C_i, C_j) \in G$. Graph G is connected if there exists a path containing edges $(C_i, C_{m_1}), (C_{m_1}, C_{m_2}), ..., (C_{m_{n-1}}, C_{m_n}), (C_{m_n}, C_j)$ in G for any C_i and C_j .

Agglomerative Merging: After the graph is established, some handcrafted features need to be extracted to compute the distance between every two nodes. These distances will then be used as criteria for merging adjacent nodes. Several types of distance are

TABLE 1: Descriptions of different cloud formations in the CCSN dataset.

Cloud Level	Cloud Genus	Abbreviation	Characteristics	Number of Images
	Cirrus	Ci	Fibrous, thin, white and transparent clouds	139
High level	Cirrocumulus	Cc	Small and white flakes arranged in groups	268
	Cirrostratus	Cs	Thin and translucent ice crystals	287
Mid level	Altocumulus	Ac	Thicker and gray line-arranged cloud sheets	221
	Altostratus	As	Opaque striped veil of grayish cloud	188
Low level	Stratus	St	Ragged and stratiform clouds that lay evenly	202
	Stratocumulus	Sc.	Dark gray layered clouds	340
	Nimbostratus	Ns	Deep gray and fluffy rain clouds	274
Vertical level	Cumulus	Cu	grayish clouds with clear contours, flat bases and circular tops	182
	Cumulonimbus	Cb	Dark-gray rain clouds with blurry and doomed edges	242

Fig. 3: Painting dataset distributions. John Constable and Boudin's paintings have the highest percentages in the dataset. Cumulus and cumulonimbus are the two most dominant cloud types.

exploited in the A3C algorithm.

(1) Color. Let μ_i and μ_j be the average LUV color vectors in clusters C_i and C_j . $\left\| \cdot \right\|_2$ denote the Euclidean distance, and n_i , n_j be the number of pixels in the patches i and j, respectively. The color distance $d_c(i, j)$ is defined as:

$$
d_c(i,j) = ||\mu_i - \mu_j||^2 \frac{n_i n_j}{n_i + n_j}.
$$
 (1)

(2) Edge. Two Sobel filters are applied to obtain the horizontal $\sqrt{g_x^2 + g_y^2}$, and the combined gradient of three color channels for and vertical derivatives g_x and g_y . The gradient is calculated by each pixel is $g = (g_l + g_u + g_v)/3$. Let b_{ij} be the boundary pixel set, then the edge distance $d_e(i, j)$ is defined as:

$$
d_e(i,j) = \frac{1}{|b_{ij}|} \sum_{k \in b_{ij}} g_k.
$$
 (2)

(3) Location. Same as the color feature, We define the Euclidean distance d_l between the average coordinates of each patch as:

$$
d_l(i,j) = ||z_i - z_j||^2 \frac{n_i n_j}{n_i + n_j},
$$
\n(3)

where z_i and z_j are average horizontal and vertical coordinates of patches i and j , respectively.

Then for patches i and j , their pairwise distance is defined as:

$$
d(i,j) = \sqrt{\lambda_c d_c(i,j)^2 + \lambda_l d_l(i,j)^2} + \lambda_e d_e(i,j).
$$
 (4)

This distance is used to merge patches that are connected with a pre-set threshold ϵ . The merging is from the patch in the smallest size at each iteration. We merge connected nodes C_i and C_j into a new node if $d(i, j) < \epsilon$. The pairwise distance will be computed iteratively after the graph is updated through the merging operation. Once no two more patches can be merged, the first-stage clustering is ended with visually similar patches. Then in the second-stage merging, we still follow the same merging strategy but incorporate the balanced partition measure and jaggedness measure [6] into the pairwise distance to achieve a better overall segmentation result. We refer to the generated segmented regions at the final state as segments.

Sky-versus-Land Classification: After obtaining these segments, we need to classify whether each segment belongs to the sky or land regions. To separate the sky and land or other irrelevant objects accurately, we perform a logistic regression for this twoclass segmentation problem. For each segment, we need to extract some features to describe these two distinct regions. Through experiments, we notice that location and color-based features can have significant impacts on the regression performance. Thus, We collected a 10-dimensional feature vector for each segment, which contains: *normalized intensity, normalized saturation, normalized hue, the square of intensity, the square of saturation, the cosine of the average hue, average vertical position, top-most vertical position, bottom-most vertical position, and the ratio between width/height by bounding box*. These features are used for regression to decide whether the segment is one of the two classes, sky or land.

In addition, we show some more sky region segmentation results in Figs. 7 and 8. After obtaining the sky regions, we compute

Tree at Right, 1821 1821 1821

Cloud Study, Hampstead, *Cloud Study: Stormy Sunset, Clouds Study*, 1822 *Study of Sky and Trees*,

of Gaius Cestius, Rome White Clouds Trees and Buildings

(a) John Constable

(b) Valenciennes

(c) Watts

Landscape with the Pyramid Rome: Study of a Cloudy Sky At the Villa Borghese: At Villa Borghese:

Strand-on-the-Green, London View of Barges on the View of the Thames from An English River, *Thames with Henley-on- Tilehurst* circa 1830-1870 *Thames Beyond*, 1830

View near Crediton, Devon, *View near Crediton, Devon Scottish Loch with Le Pont du Gard*

1867 *Game Birds*, 1852

Fig. 4: Representative paintings of the seven artists in our dataset. The figure is continued in Fig. 5.

(d) Lee

the hue distribution of each painting collection by converting the color space to HSV. Fig. 9 shows the hue distribution by counting the number of pixels in the sky region belonging to each hue value (0-360) and the Kernel Density Estimation (KDE) of each distribution.

4 CLOUD CLASSIFICATION

Our neural network contains two parts, a pre-learned edge feature encoder and a classic feature encoder. The pipeline is shown in Fig. 10. An exemplary output of the edge feature encoder is shown in Fig. 11 (d).

To find what features are most important for the CNN classification model, we use the Grad-cam visualization method [7], which provides a heatmap indicating the significance of any location in the feature map for reaching the classification decision. In Figs. 11 (b) and (c), the visualization result for an example image based on the final convolution layer in the last Resconv module shows that the edge information of each cloud mass is important for classifying the cloud type. We are thus motivated to directly include edge- or contour-related features in the neural network to increase classification accuracy.

A schematic plot for the extraction of content and style features by MUNIT [8] is given in Fig. 12.

5 MULTIDIMENSIONAL SCALING RESULTS

To better understand the style distances between individual paintings in the entire collection, we generate two plots to show the

Dedham Water Meadows View in Kent View of Hampstead looking Beach near Yarmouth, *towards Harrow*, *circa* 1850 *towards Harrow,* circa 1860-1880 (e) Lionel Constable

 S *alute from San Giorgio*, 1895

Venice, Santa Maria della Harbor Scene Beaulieu: The Bay of Fourmis, *Port of Le Havre*, 1886

Moorland Road, 1851 *A Windy Day*, 1850 *The Road across the Rhyl Sands*, 1854

Common, 1853

Fig. 5: Representative paintings of the seven artists in our dataset. Continued from Fig. 4.

(g) Cox

(f) Stratus (g) Stratocumulus (h) Nimbostratus (i) Cumulus (j) Cumulonimbus

Fig. 6: Representative photographs of different types of clouds in the CCSN dataset.

(a) *Road to the Spaniards, Hampstead*, John Constable, 1822

(b) *Rome: Study of a cloudy sky*, Valenciennes

(c) *View of the Thames from Tilehurst*, Watts

(e) *Landscape with Wheatfield*, Lionel Constable, circa 1850

Fig. 7: Sky and ground segmentation illustrated with a painting by each artist. Left: Original painting. Middle: Segments generated after a two-round merging. Right: Sky and land segmentation maps. This figure is continued in Fig. 8.

(f) *A Marine Scene*, Boudin, 1878

(g) Photo from the CCSN dataset

Fig. 8: Sky and ground segmentation illustrated with a painting by each artist and photos from the Middle: Segments generated after a two-round merging. Right: Sky and land segmentation maps. Continued from Fig. 7.

Fig. 9: The hue distribution of all seven artists' painting collections and their corresponding kernel density estimate results.

Fig. 10: The structure of our cloud classification. First, we need to get segmented images to use as input to the network. Then, two streams of encoders aim for extracting classic and edge features.

Fig. 11: Grad-cam visualization. (a) Example cloud painting. (b) The Grad-cam heatmap highlights where the model relies on the most to conclude the class of the image. Warmer colors indicate a higher significance of a location in the feature map. Red is the warmest, with yellow, green, blue, and purple becoming increasingly cooler. (c) The guided back-propagation plot is another way to show the contribution of features to the classification result. Brighter pixels indicate that the features at their positions are more important. (d) The output edge estimation of the HED model.

Fig. 12: The process of image translation from paintings to photos with content-style disentanglement.

multidimensional scaling (MDS) results of these paintings using the style distances between any pair of paintings applied to groups each containing a single painting). Figs. 13 and 14 show the MDS results in two dimensions.

Fig. 13: Multidimensional scaling (MDS) results of paired paintings. The MDS plot shows all 363 paintings. Only the sky regions are used in the analysis.

6 STYLE DISTANCE IN THE EXPANDED DATASET

Besides the seven painters discussed so far, we expanded our dataset to include landscape paintings by artists working in diverse styles from the Renaissance painter Titian (c. 1490-1576), to the 20th-century modernists Georgia O'Keeffe (1887-1986) and Thomas Hart Benton (1889-1975), and the popular contemporary landscapist Thomas Kinkade (1958-2012) as well as watercolors by John Constable to show that the proposed style distance can be applied to more artists and media. The style distances between these artists and John Constable are provided in Table 2. The statistics of T-test about whether these painting collections are similar to John Constable's are listed in Table 3.

7 STYLE SIMILARITY OF WHOLE PAINTINGS

Instead of using only the sky regions to analyze the style transfer, we also obtained the trained style encoder using the whole images of both the paintings and photographs for training to measure the similarity to photos and style distance among the whole paintings of each artist. We utilized the same pipeline and settings in Section TABLE 2: Style distance among different painting collections.

Pair of Painting Collections in Comparison	D_{style} (mean \pm std)
(John Constable:oil, John Constable:watercolor)	0.606 ± 0.143
(John Constable, Benton)	0.835 ± 0.149
(John Constable, Titian)	0.843 ± 0.138
(John Constable, Kinkade)	0.881 ± 0.156
(John Constable, O'Keeffe)	0.926 ± 0.153

TABLE 3: T statistics of T -test about the difference of D_{style}

5.4 to sample data and conduct hypothesis testing but using the style features generated from the style encoder trained with the

Fig. 14: Multidimensional scaling results of paired paintings. The colors of the scattered points indicate different painters.

whole paintings. The corresponding R_{style} and the statistics of T test about whether John Constable's paintings are more similar to photos are listed in Tables 4 and 5. It is worth noting that John Constable's paintings are not significantly more similar to photographs than those painted by his son Lionel Constable at the significance level of 0.1, while do have a smaller style distance to photographs than other painters. In addition, D_{style} and the statistics of T-test about whether the other six artists' painting styles are similar to John Constable's are listed in Tables 6 and 7. We can still observe that John Constable and Lionel Constable shared a similar painting style at the significance level of 0.01. The Multidimensional scaling (MDS) results of D_{style} computed using the whole paintings are shown in Figs. 15 and 16.

TABLE 4: R_{style} of the painting collection of each painter using the whole painting. These R_{style} are generated with the same sampling method, but using the whole images of both paintings and photos to train the style encoder.

Artist	R_{style} (mean \pm std)	
John Constable	1.381 ± 0.169	
Lionel Constable	1.403 ± 0.162	
Valenciennes	1.492 ± 0.178	
Watts	1.515 ± 0.176	
Lee	1.548 ± 0.181	
Boudin	1.562 ± 0.183	
Cox	1.581 ± 0.179	

Fig. 15: Multidimensional scaling (MDS) results of paired paintings using the style distances between any pair of paintings. The style features are generated using the style encoder trained with the whole images. The MDS plot shows all 363 paintings.

TABLE 5: T statistics of T -test about the difference of R_{style} between John Constable and other artists using the whole images. Based on these statistics, John constable's painting style of landscape paintings is more similar to real-world scenes than all other artists except Lionel Constable at significance level 0.1.

TABLE 6: Style distance among different painting collections. These D_{style} 's are generated with the same sampling method, but using the whole images of both paintings and photos to train the style encoder.

Fig. 16: MDS results of paired paintings using the style distances between any pair of paintings. The style features are generated using the style encoder trained with the whole images. The colors of the scattered points indicate different painters.

TABLE 7: T statistics of T -test about the difference of D_{style} using the whole image. Based on these statistics, John Constable's painting style of landscape paintings is similar to Lionel Constable's at the significance level of 0.01.

Artist	T -statistic	p -value
Lionel Constable	-2.120	0.018
Valenciennes	-3.712	$1.27e-4$
Boudin	-4.087	2.980e-05
Cox	-5.416	7.380e-08
Watts	-6.296	$7.300e-10$
Lee	-7.291	2.339e-12

REFERENCES

- [1] L. Howard, "On the modifications of clouds, and on the principles of their production, suspension, and destruction: Being the substance of an essay read before the Askesian Society in the session 1802–3," *The Philosophical Magazine*, pp. 5–11, 1803.
- [2] J. Zhang, P. Liu, F. Zhang, and Q. Song, "CloudNet: Ground-based cloud classification with deep convolutional neural network," *Geophysical Research Letters*, vol. 45, no. 16, pp. 8665–8672, 2018.
- [3] R. P. Mihail, S. Workman, Z. Bessinger, and N. Jacobs, "Sky segmentation in the wild: An empirical study," in *2016 IEEE Winter Conf. Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–6.
- [4] S. Rawat, S. Gairola, R. Shah, and P. Narayanan, "Find me a sky: A datadriven method for color-consistent sky search and replacement," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2018, pp. 216–228.
- [5] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 633–641.
- [6] J. Li, "Agglomerative connectivity constrained clustering for image segmentation," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 1, pp. 84–99, 2011.
- [7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradientbased localization," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 618–626.
- [8] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 172–189.

TABLE 8: Paintings in our dataset and their main cloud and weather types as determined by an expert meteorologist. Instead of only providing cloud types for each painting, the detailed cloud structure, corresponding painting environments and the evaluation of realism are also offered. 'NG' denotes 'Not given'.

