Recognizing Image Style

Sergey Karayev[1]

Matthew Trentacoste[2]

Helen Han[1]

Aseem Agarwala[2]

Trevor Darrell[1]

Aaron Hertzmann[2]

Holger Winnemoeller[2]

[1] University of California, Berkeley

[2] Adobe

arXiv:1311.3715v3 [cs.CV] 23 Jul 2014

Abstract

The style of an image plays a significant role in how it is viewed, but style has received little attention in computer vision research. We describe an approach to predicting style of images, and perform a thorough evaluation of different image features for these tasks. We find that features learned in a multi-layer network generally perform best – even when trained with object class (not style) labels. Our large-scale learning methods results in the best published performance on an existing dataset of aesthetic ratings and photographic style annotations. We present two novel datasets: 80K Flickr photographs annotated with 20 curated style labels, and 85K paintings annotated with 25 style/genre labels. Our approach shows excellent classification performance on both datasets. We use the learned classifiers to extend traditional tag-based image search to consider stylistic constraints, and demonstrate cross-dataset understanding of style.

1 Introduction

Deliberately-created images convey meaning, and *visual style* is often a significant component of image meaning. For example, a political candidate portrait made in the lush colors of a Renoir painting tells a different story than if it were in the harsh, dark tones of a horror movie. Distinct visual styles are apparent in art, cinematography, advertising, and have become extremely popular in amateur photography, with apps like Instagram leading the way. While understanding style is crucial to image understanding, very little research in computer vision has explored visual style.

Although is it very recognizable to human observers, visual style is a difficult concept to rigorously define. Most academic discussion of style has been in an art history context, but the distinctions between, say, Rococo versus pre-Rafaelite style are less relevant to modern photography and design. There has been some previous research in image style, but this has principally been limited to recognizing a few, well-defined optical properties, such as depth-of-field.

We define several different *types* of image style, and gather a new, large-scale dataset of photographs annotated with style labels. This dataset embodies several different aspects
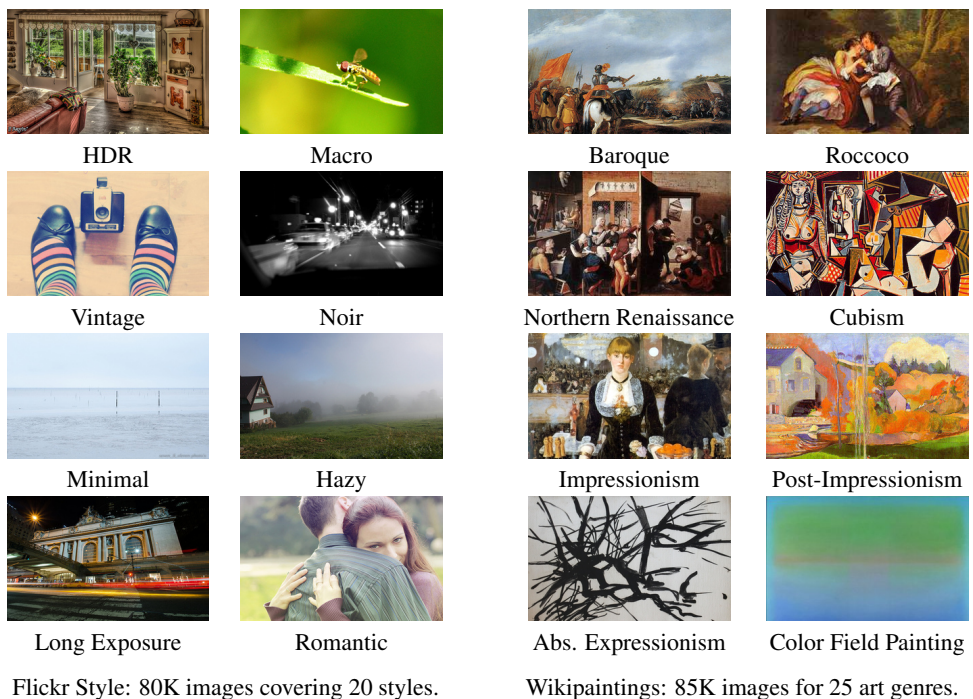
HDR	Macro	Baroque	Roccoco
Vintage	Noir	Northern Renaissance	Cubism
Minimal	Hazy	Impressionism	Post-Impressionism
Long Exposure	Romantic	Abs. Expressionism	Color Field Painting

Flickr Style: 80K images covering 20 styles.　　Wikipaintings: 85K images for 25 art genres.

Figure 1: Typical images in different style categories of our datasets.

of visual style, including photographic techniques ("Macro," "HDR"), composition styles ("Minimal," "Geometric"), moods ("Serene," "Melancholy"), genres ("Vintage," "Romantic," "Horror"), and types of scenes ("Hazy," "Sunny"). These styles are not mutually exclusive, and represent different attributes of style. We also gather a large dataset of visual art (mostly paintings) annotated with art historical style labels, ranging from Renaissance to modern art. Figure 1 shows some samples.

We test existing classification algorithms on these styles, evaluating several state-of-the-art image features. Most previous work in aesthetic style analysis has used hand-tuned features, such as color histograms. We find that deep convolutional neural network (CNN) features perform best for the task. This is surprising for several reasons: these features were trained on object class categories (ImageNet), and many styles appear to be primarily about color choices, yet the CNN features handily beat color histogram features. This leads to one conclusion of our work: mid-level features derived from object datasets are generic for style recognition, and superior to hand-tuned features.

We compare our predictors to human observers, using Amazon Mechanical Turk experiments, and find that our classifiers predict Group membership at essentially the same level of accuracy as Turkers. We also test on the AVA aesthetic prediction task [22], and show that using the "deep" object recognition features improves over the state-of-the-art results.

Applications and code.　First, we demonstrate an example of using our method to search for images by style. This could be useful for applications such as product search, storytelling, and creating slide presentations. In the same vein, visual similarity search results could be

filtered by visual style, making possible queries such as "similar to this image, but more Film Noir." Second, style tags may provide valuable mid-level features for other image understanding tasks. For example, there has increasing recent effort in understanding image meaning, aesthetics, interestingness, popularity, and emotion (for example, [10, 12, 14, 16]), and style is an important part of meaning. Finally, learned predictors could be a useful component in modifying the style of an image.

All data, trained predictors, and code (including results viewing interface) are available at http://sergeykarayev.com/recognizing-image-style/.

2 Related Work

Most research in computer vision addresses recognition and reconstruction, independent of image style. A few previous works have focused directly on image composition, particularly on the high-level attributes of beauty, interestingness, and memorability.

Most commonly, several previous authors have described methods to predict aesthetic quality of photographs. Datta et al. [4], designed visual features to represent concepts such as colorfulness, saturation, rule-of-thirds, and depth-of-field, and evaluated aesthetic rating predictions on photographs; The same approach was further applied to a small set of Impressionist paintings [18]. The feature space was expanded with more high-level descriptive features such as "presence of animals" and "opposing colors" by Dhar et al., who also attempted to predict Flickr's proprietary "interestingness" measure, which is determined by social activity on the website [6]. Gygli et al. [10] gathered and predicted human evaluation of image interestingness, building on work by Isola et al. [12], who used various high-level features to predict human judgements of image memorability. In a similar task, Borth et al. [3] performed sentiment analysis on images using object classifiers trained on adjective-noun pairs.

Murray et al. [22] introduced the Aesthetic Visual Analysis (AVA) dataset, annotated with ratings by users of DPChallenge, a photographic skill competition website. The AVA dataset contains some photographic style labels (e.g., "Duotones," "HDR"), derived from the titles and descriptions of the photographic challenges to which photos were submitted. Using images from this dataset, Marchesotti and Peronnin [20] gathered bi-grams from user comments on the website, and used a simple sparse feature selection method to find ones predictive of aesthetic rating. The attributes they found to be informative (e.g., "lovely photo," "nice detail") are not specific to image style.

Several previous authors have developed systems to classify classic painting styles, including [15, 25]. These works consider only a handful of styles (less than ten apiece), with styles that are visually very distinct, e.g., Pollock vs. Dalí. These datasets comprise less than 60 images per style, for both testing and training. Mensink [21] provides a larger dataset of artworks, but does not consider style classification.

3 Data Sources

Building an effective model of photographic style requires annotated training data. To our knowledge, there is only one existing dataset annotated with visual style, and only a narrow range of photographic styles is represented [22]. We would like to study a broader range of styles, including different *types* of styles ranging from genres, compositional styles, and

moods. Morever, large datasets are desirable in order to obtain effective results, and so we would like to obtain data from online communities, such as Flickr.

Flickr Style. Although Flickr users often provide free-form tags for their uploaded images, the tags tend to be quite unreliable. Instead, we turn to Flickr groups, which are community-curated collections of visual concepts. For example, the Flickr Group "Geometry Beauty" is described, in part, as "Circles, triangles, rectangles, symmetric objects, repeated patterns", and contains over 167K images at time of writing; the "Film Noir Mood" group is described as "Not just black and white photography, but a dark, gritty, moody feel..." and comprises over 7K images.

At the outset, we decided on a set of 20 visual styles, further categorized into types:

- **Optical techniques:** Macro, Bokeh, Depth-of-Field, Long Exposure, HDR
- **Atmosphere:** Hazy, Sunny
- **Mood:** Serene, Melancholy, Ethereal
- **Composition styles:** Minimal, Geometric, Detailed, Texture
- **Color:** Pastel, Bright
- **Genre:** Noir, Vintage, Romantic, Horror

For each of these stylistic concepts, we found at least one dedicated Flickr Group with clearly defined membership rules. From these groups, we collected 4,000 positive examples for each label, for a total of 80,000 images. Example images are shown in Figure 1a. The exact Flickr groups used are given in Table 2.

The derived labels are considered clean in the positive examples, but may be noisy in the negative examples, in the same way as the ImageNet dataset [5]. That is, a picture labeled as *Sunny* is indeed *Sunny*, but it may also be *Romantic*, for which it is not labeled. We consider this an unfortunate but acceptable reality of working with a large-scale dataset. Following ImageNet, we still treat the absence of a label as indication that the image is a negative example for that label. Mechanical Turk experiments described in section 6.1 serve to allay our concerns.

Wikipaintings. We also provide a new dataset for classifying painting style. To our knowledge, no previous large-scale dataset exists for this task – although very recently a large dataset of artwork did appear for other tasks [21]. We collect a dataset of 100,000 high-art images – mostly paintings – labeled with artist, style, genre, date, and free-form tag information by a community of experts on the `Wikipaintings.org` website.

Analyzing style of non-photorealistic media is an interesting problem, as much of our present understanding of visual style arises out of thousands of years of developments in fine art, marked by distinct historical styles. Our dataset presents significant stylistic diversity, primarily spanning Renaissance styles to modern art movements (Figure 6 provides further breakdowns). We select 25 styles with more than 1,000 examples, for a total of 85,000 images. Example images are shown in Figure 1b.

4 Learning algorithm

We learn to classify novel images according to their style, using the labels assembled in the previous section. Because the datasets we deal with are quite large and some of the features

are high-dimensional, we consider only linear classifiers, relying on sophisticated features to provide robustness.

We use an open-source implementation of Stochastic Gradient Descent with adaptive subgradient [8]. The learning process optimizes the function

$$\min_w \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} \|w\|_2^2 + \sum_i \ell(x_i, y_i, w)$$

We set the L_1 and L_2 regularization parameters and the form of the loss function by validation on a held-out set. For the loss $\ell(x, y, w)$, we consider the hinge $(\max(0, 1 - y \cdot w^T x))$ and logistic $(\log(1 + \exp(-y \cdot w^T x)))$ functions. We set the initial learning rate to 0.5, and use adaptive subgradient optimization [8]. Our setup is of multi-class classification; we use the One vs. All reduction to binary classifiers.

5 Image Features

In order to classify styles, we must choose appropriate image features. We hypothesize that image style may be related to many different features, including low-level statistics [19], color choices, composition, and content. Hence, we test features that embody these different elements, including features from the object recognition literature. We evaluate single-feature performance, as well as second-stage fusion of multiple features.

L*a*b color histogram. Many of the Flickr styles exhibit strong dependence on color. For example, *Noir* images are nearly all black-and-white, while most *Horror* images are very dark, and *Vintage* images use old photographic colors. We use a standard color histogram feature, computed on the whole image. The 784-dimensional joint histogram in CIELAB color space has 4, 14, and 14 bins in the L*, a*, and b* channels, following Palermo et al. [24], who showed this to be the best performing single feature for determining the date of historical color images.

GIST. The classic gist descriptor [23] is known to perform well for scene classification and retrieval of images visually similar at a low-resolution scale, and thus can represent image composition to some extent. We use the INRIA LEAR implementation, resizing images to 256 by 256 pixels and extracting a 960-dimensional color GIST feature.

Graph-based visual saliency. We also model composition with a visual attention feature [11]. The feature is fast to compute and has been shown to predict human fixations in natural images basically as well as an individual human (humans are far better in aggregate, however). The 1024-dimensional feature is computed from images resized to 256 by 256 pixels.

Meta-class binary features. Image content can be predictive of individual styles, e.g., *Macro* images include many images of insects and flowers. The `mc-bit` feature [2] is a 15,000-dimensional bit vector feature learned as a non-linear combination of classifiers trained using existing features (e.g., SIFT, GIST, Self-Similarity) on thousands of random ImageNet synsets, including internal ILSVRC2010 nodes. In essence, MC-bit is a hand-crafted "deep" architecture, stacking classifiers and pooling operations on top of lower-level features.

Table 1: Mean APs on three datasets for the considered single-channel features and their second-stage combination. As some features were clearly worse than others on the AVA Style dataset, only the better features were evaluated on larger datasets.

	Fusion x Content	DeCAF$_6$	MC-bit	L*a*b* Hist	GIST	Saliency	random
AVA Style	**0.581**	0.579	0.539	0.288	0.220	0.152	0.132
Flickr	**0.368**	0.336	0.328	-	-	-	0.052
Wikipaintings	**0.473**	0.356	0.441	-	-	-	0.043

Deep convolutional net. Current state-of-the-art results on ImageNet, the largest image classification challenge, have come from a deep convolutional network trained in a fully-supervised manner [12]. We use the Caffe [13] open-source implementation of the ImageNet-winning eght-layer convolutional network, trained on over a million images annotated with 1,000 ImageNet classes. We investigate using features from two different levels of the network, referred to as DeCAF$_5$ and DeCAF$_6$ (following [7]). The features are 8,000- and 4,000-dimensional and are computed from images center-cropped and resized to 256 by 256 pixels.

Content classifiers. Following Dhar et al. [6], who use high-level classifiers as features for their aesthetic rating prediction task, we evaluate using object classifier confidences as features. Specifically, we train classifiers for all 20 classes of the PASCAL VOC [9] using the DeCAF$_6$ feature. The resulting classifiers are quite reliable, obtaining 0.7 mean AP on the VOC 2012.

We aggregate the data to train four classifiers for "animals", "vehicles", "indoor objects" and "people". These aggregate classes are presumed to discriminate between vastly different types of images – types for which different style signals may apply. For example, a *Romantic* scene with people may be largely about the composition of the scene, whereas, *Romantic* scenes with vehicles may be largely described by color.

To enable our classifiers to learn content-dependent style, we can take the outer product of a feature channel with the four aggregate content classifiers.

6 Experiments

6.1 Flickr Style

We learn and predict style labels on the 80,000 images labeled with 20 different visual styles of our new Flickr Style dataset, using 20% of the data for testing, and another 20% for parameter-tuning validation.

There are several performance metrics we consider. Average Precision evaluation (as reported in Table 1 and in Table 4) is computed on a random class-balanced subset of the test data (each class has equal prevalence). We compute confusion matrices (Figure 8, Figure 9, Figure 7) on the same data. Per-class accuracies are computed on subsets of the data balanced by the binary label, such that chance performance is 50%. We follow these decisions in all following experiments.

The best single-channel feature is DeCAF$_6$ with 0.336 mean AP; feature fusion obtains 0.368 mean AP. Per-class APs range from 0.17 [Depth of Field] to 0.62 [Macro]. Per-class accuracies range from 68% [Romantic, Depth of Field] to 85% [Sunny, Noir, Macro]. The

average per-class accuracy is 78%. We show the most confident style classifications on the test set of Flickr Style in Figure 3.

Upon inspection of the confusion matrices, we saw points of understandable confusion: Depth of Field vs. Macro, Romantic vs. Pastel, Vintage vs. Melancholy. There are also surprising sources of mistakes: Macro vs. Bright/Energetic, for example. To explain this particular confusion, we observed that lots of Macro photos contain bright flowers, insects, or birds, often against vibrant greenery. Here, at least, the content of the image dominates its style label.

To explore further content-style correlations, we plot the outputs of PASCAL object class classifiers (one of our features) on the Flickr dataset in Figure 2. We can observe that some styles have strong correlations to content (e.g., "Hazy" occurs with "vehicle", "HDR" doesn't occur with "cat").

We hypothesize that style is content-dependent: a Romantic portrait may have different low-level properties than a Romantic sunset. We form a new feature as an outer product of our content classifier features with the second-stage late fusion features ("Fusion × Content" in all results figures). These features gave the best results, thus supporting the hypothesis.

	Detailed	Pastel	Melancholy	Noir	HDR	Vintage	Long Exposure	Horror	Sunny	Bright	Hazy	Bokeh	Serene	Texture	Ethereal	Macro	Depth of Field	Geometric	Minimal	Romantic
animal	0.04	-0.00	-0.02	-0.06	-0.05	-0.04	-0.04	-0.01	-0.07	-0.02	-0.05	0.11	0.01	0.03	0.02	0.22	0.06	-0.06	-0.04	-0.03
indoor	0.07	0.05	-0.06	-0.01	-0.05	0.04	-0.06	-0.04	-0.10	0.06	-0.10	0.03	-0.06	0.05	-0.06	0.07	0.00	0.11	0.05	0.00
person	-0.05	0.06	0.10	0.09	-0.04	0.07	-0.07	0.14	-0.07	-0.00	-0.05	0.02	-0.05	-0.06	0.02	-0.10	0.04	-0.07	-0.09	0.10
vehicle	-0.00	-0.07	-0.04	-0.03	0.12	-0.05	0.17	-0.08	0.18	-0.00	0.11	-0.07	0.06	-0.05	-0.06	-0.08	-0.05	0.01	-0.00	-0.04

Figure 2: Correlation of PASCAL content classifier predictions (rows) against ground truth Flickr Style labels (columns). We see, for instance, that the Macro style is highly correlated with presence of animals, and that Long Exposure and Sunny style photographs often feature vehicles.

Mechanical Turk Evaluation. In order to provide a human baseline for evaluation, we performed a Mechanical Turk study. For each style, Turkers were shown positive and negative examples for each Flickr Group, and then they evaluated whether each image in the test set was part of the given style. We treat the Flickr group memberships as ground truth as before, and then evaluate Turkers' ability to accurately determine group membership. Measures were taken to remove spam workers; see **??** for our experimental setup. For efficiency, one quarter of the test set was used, and two redundant styles (Bokeh and Detailed) were removed. Each test image was evaluated by 3 Turkers, and the majority vote taken as the human result for this image.

Results are presented in Table 6. In total, Turkers achieved 75% mean accuracy (ranging from 61% [Romantic] to 92% [Macro]) across styles, in comparison to 78% mean accuracy (ranging from 68% [Depth of Field] to 87% [Macro]) of our best method. Our algorithm did significantly worse than Turkers on Macro and Horror, and significantly better on Vintage, Romantic, Pastel, Detailed, HDR, and Long Exposure styles.

Some of this variance may be due to subtle difference from the Turk tasks that we provided, as compared to the definitions of the Flickr groups, but may also due to the Flickr groups' incorporating images that do not quite fit the common definition of the given style. For example, there may be a mismatch between different notions of "romantic" and "vintage," and how inclusively these terms are defined.

We additionally used the Turker opinion as ground truth for our method's predictions. In switching from the default Flickr to the MTurk ground truth, our method's accuracy hardly changed from 78% to 77%. However, we saw that the accuracy of our Vintage, Detailed, Long Exposure, Minimal, HDR, and Sunny style classifiers significantly decreased, indicating machine-human disagreement on those styles.

6.2   Wikipaintings

With the same setup and features as in the Flickr experiments, we evaluate 85,000 images labeled with 25 different art styles. Detailed results are provided in Table 5 and Table 7. The best single-channel feature is MC-bit with 0.441 mean AP; feature fusion obtains 0.473 mean AP. Per-class accuracies range from 72% [Symbolism, Expressionism, Art Nouveau] to 94% [Ukiyo-e, Minimalism, Color Field Painting].

6.3   AVA Style

AVA [22] is a dataset of 250K images from `dpchallenge.net`. We evaluate classification of aesthetic rating and of 14 different photographic style labels on the 14,000 images of the AVA dataset that have such labels. For the style labels, the publishers of the dataset provide a train/test split, where training images have only one label, but test images may have more than one label [22]. Our results are presented in Table 3. For style classification, the best single feature is the $DeCAF_6$ convolution network feature, obtaining 0.579 mean AP. Feature fusion improves the result to 0.581 mean AP; both results beat the previous state-of-the-art of 0.538 mean AP [22]. [1]

In all metrics, the DeCAF and MC-bit features significantly outperformed more low-level features on this dataset. Accordingly, we do not evaluate the low-level features on the larger Flickr and Wikipaintings datasets.

Test images were grouped into 10 images per Human Interface Task (HIT). Each task asks the Turker to evaluate the style (e.g., "Is this image VINTAGE?") for each image. For each style, we provided a short blurb describing the style in words, and provided 12-15 hand-chosen positive and negative examples for each Flickr Group. Each HIT included 2 sentinels: images which were very clearly positives and similar to the examples. HITs were rejected when Turkers got both sentinels wrong. Turkers were paid 0.10 per HIT, and were allowed to perform multiple hits. Manual inspection of the results indicate that the Turkers understood the task and were performing effectively. A few Turkers sent unsolicited feedback indicating that they were really enjoying the HITs ("some of the photos are beautiful") and wanted to perform them as effectively as possible.
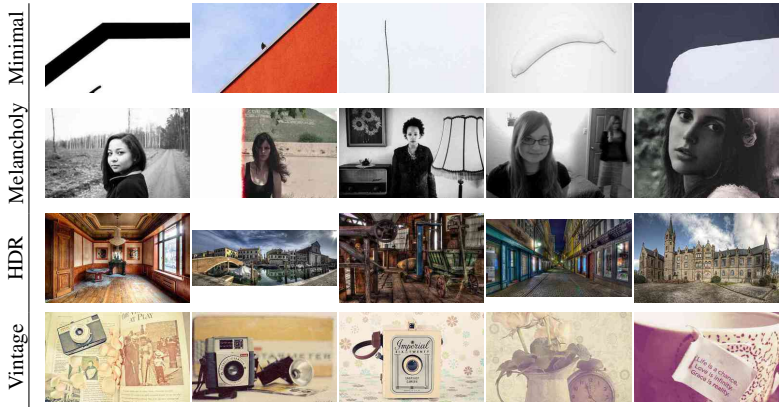
Figure 3: Top five most-confident positive predictions on the Flickr Style test set, for a few different styles.

6.4 Application: Style-Based Image Search

Style classifiers learned on our datasets can be used toward novel goals. For example, sources of stock photography or design inspiration may be better navigated with a vocabulary of style. Currently, companies expend labor to manually annotate stock photography with such labels. With our approach, any image collection can be searchable and rankable by style.

To demonstrate, we apply our Flickr-learned style classifiers to a new dataset of 80K images gathered on Pinterest (also available with our code release); some results are shown in Figure 5. Interestingly, styles learned from photographs can be used to order paintings, and styles learned from paintings can be used to order photographs, as illustrated in Figure 4.

Figure 4: Cross-dataset style. On the left are shown top scorers from the Wikipaintings set, for styles learned on the Flickr set. On the right, Flickr photographs are accordingly sorted by Painting style. (Figure best viewed in color.)

---

[1]Our results beat 0.54 mAP using both the AVA-provided class-imbalanced test split, and the class-balanced subsample that we consider to be more correct evaluation, and for which we provide numbers.

Query: "dress".

Query: "flower".
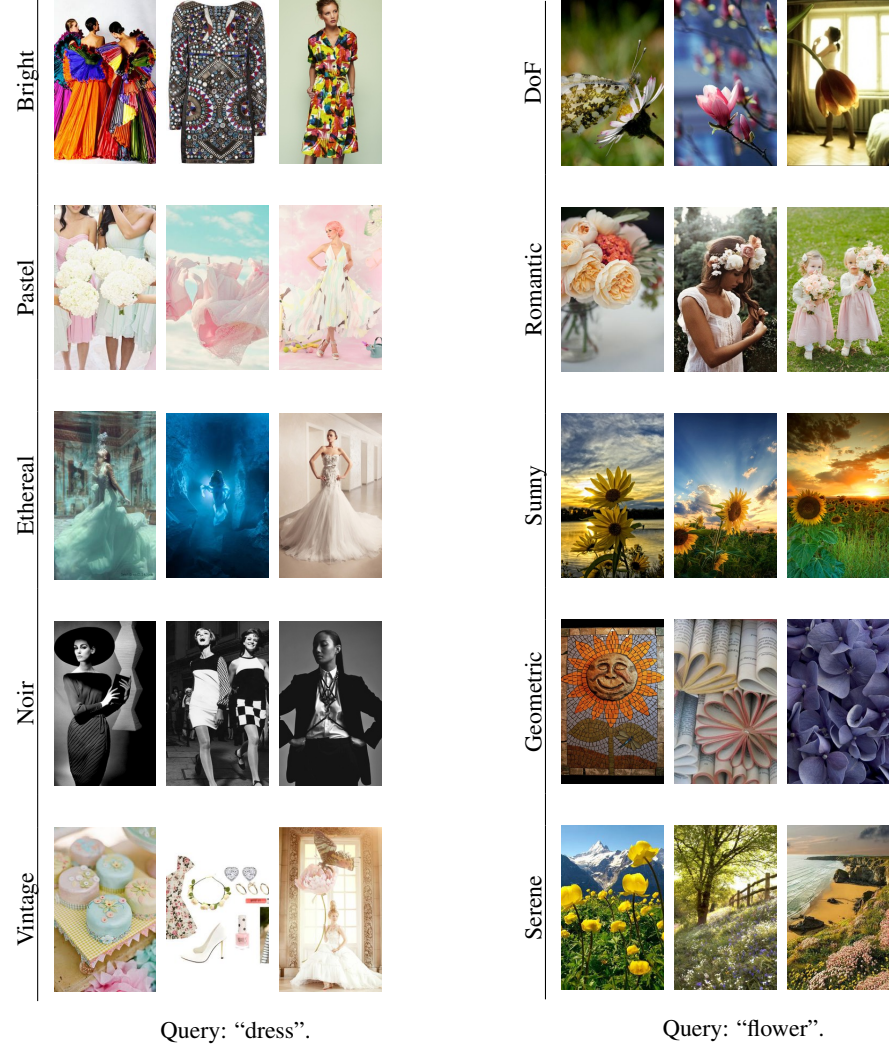
Figure 5: Example of filtering image search results by style. Our Flickr Style classifiers are applied to images found on Pinterest. The images are searched by the text contents of their captions, then filtered by the response of the style classifiers. Here we show three out of top five results for different query/style combinations.

6.5 Discussion

We have made significant progress in defining the problem of understanding photographic style. We provide a novel dataset that exhibits several types of styles not previously considered in the literature, and we demonstrate state-of-the-art results in prediction of both style and aesthetic quality. These results are comparable to human performance. We also show that style is highly content-dependent.

Style plays a significant role in much of the manmade imagery we experience daily, and there is considering need for future work to further answer the question "What is style?"

One of the most interesting outcomes of this work is the success of features trained for object detection for both aesthetic and style classification. We propose several possible hypotheses to explain these results. Perhaps the network layers that we use as features are extremely good as general visual features for image representation in general. Another explanation is that object recognition depends on object appearance, e.g., distinguishing red from white wine, or different kinds of terriers, and that the model learns to repurpose these features for image style. Understanding and improving on these results is fertile ground for future work.

References

[1] Alekh Agarwal, Olivier Chapelle, Miroslav Dudik, and John Langford. A Reliable Effective Terascale Linear Learning System. *Journal of Machine Learning Research*, 2012.

[2] A. Bergamo and L. Torresani. Meta-class Features for Large-Scale Object Categorization on a Budget. In *CVPR*, 2012.

[3] Damian Borth, Rongrong Ji, Tao Chen, and Thomas M Breuel. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM MM*, 2013.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*, 2006.

[5] Jia Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[6] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR*, 2011.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, Trevor Darrell, Trevor Eecs, and Berkeley Edu. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. Technical report, 2013. arXiv:1310.1531v1.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 2011.

[9] M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. The PASCAL VOC Challenge Results, 2010.

[10] Michael Gygli, Fabian Nater, and Luc Van Gool. The Interestingness of Images. In *ICCV*, 2013.

[11] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-Based Visual Saliency. In *NIPS*, 2006.

[12] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What Makes an Image Memorable? In *CVPR*, 2011.

[13] Yangqing Jia. Caffe: an open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[14] Jungseock Joo, Weixin Li, Francis Steen, and Song-Chun Zhu. Visual Persuasion: Inferring Communicative Intents of Images. In *CVPR*, 2014.

[15] Daniel Keren. Painter Identification Using Local Features and Naive Bayes. In *ICPR*, 2002.

[16] A. Khosla, A. Das Sarma, and R. Hamid. What Makes an Image Popular? In *WWW*, 2014.

Table 2: Exact Flickr group names, and their sizes.

Style	Group names [num images]
Bokeh	Bokeh Photography (1/day) [187K]
Bright	Colour Mania [100K]
Depth of Field	Depth of Field [116K], Finest DoF [54K]
Detailed	Details aller Art - Details of all kind [22K], Detail pictures [5K]
Ethereal	Ethereal World [21K]
Geometric Composition	Geometric Beauty [168K]
Hazy	Misty hazy smokey [14K]
HDR	HDR ADDICTED [374K]
Horror	Horror [16K]
Long Exposure	Long Exposure [619K]
Macro	Closer and Closer Macro Photography [990K]
Melancholy	melancholy [106K]
Minimal	Less Is More... [44K]
Noir	Film Noir Mood [7K]
Romantic	Romantic Images [20K]
Serene	Šerene ̃[68K]
Pastel	pastel and dreamy [120K], Pastel Soft tone [7K]
Sunny	Sun, sun and more sun [23K]
Texture	Texture [103K]
Vintage	Vintage Feelings [4K], Vintage & Retro [61K]

[17] Alex Krizhevsky, Ilya Sutskever, and Geoff E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

[18] Congcong Li and Tsuhan Chen. Aesthetic Visual Quality Assessment of Paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009.

[19] Siwei Lyu, Daniel Rockmore, and Hany Farid. A Digital Technique for Art Authentication. *PNAS*, 101(49), 2004.

[20] Luca Marchesotti and Florent Perronnin. Learning beautiful (and ugly) attributes. In *BMVC*, 2013.

[21] Thomas Mensink and Jan van Gemert. The Rijksmuseum Challenge: Museum-Centered Visual Recognition. In *ICMR*, 2014.

[22] Naila Murray, De Barcelona, Luca Marchesotti, and Florent Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, 2012.

[23] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42(3):145–175, 2001.

[24] Frank Palermo, James Hays, and Alexei A Efros. Dating Historical Color Images. In *ECCV*, 2012.

[25] Lior Shamir, Tomasz Macura, Nikita Orlov, D. Mark Eckley, and Ilya G. Goldberg. Impressionism, Expressionism, Surrealism: Automated Recognition of Painters and Schools of Art. *ACM Trans. Applied Perc.*, 7(2), 2010.

Table 3: All per-class APs on all evaluated features on the AVA Style dataset.

	Fusion	DeCAF$_6$	MC-bit	Murray	L*a*b*	GIST	Saliency
Complementary_Colors	0.469	0.548	0.329	0.440	0.294	0.223	0.111
Duotones	0.676	0.737	0.612	0.510	0.582	0.255	0.233
HDR	0.669	0.594	0.624	0.640	0.194	0.124	0.101
Image_Grain	0.647	0.545	0.744	0.740	0.213	0.104	0.104
Light_On_White	0.908	0.915	0.802	0.730	0.867	0.704	0.172
Long_Exposure	0.453	0.431	0.420	0.430	0.232	0.159	0.147
Macro	0.478	0.427	0.413	0.500	0.230	0.269	0.161
Motion_Blur	0.478	0.467	0.458	0.400	0.117	0.114	0.122
Negative_Image	0.595	0.619	0.499	0.690	0.268	0.189	0.123
Rule_of_Thirds	0.352	0.353	0.236	0.300	0.188	0.167	0.228
Shallow_DOF	0.624	0.659	0.637	0.480	0.332	0.276	0.223
Silhouettes	0.791	0.801	0.801	0.720	0.261	0.263	0.130
Soft_Focus	0.312	0.354	0.290	0.390	0.127	0.126	0.114
Vanishing_Point	0.684	0.658	0.685	0.570	0.123	0.107	0.161
mean	0.581	0.579	0.539	0.539	0.288	0.220	0.152

Table 4: All per-class APs on all evaluated features on the Flickr dataset.

	Fusion x Content	DeCAF$_6$	MC-bit
Bokeh	0.288	0.253	0.248
Bright	0.251	0.236	0.183
Depth_of_Field	0.169	0.152	0.148
Detailed	0.337	0.277	0.278
Ethereal	0.408	0.393	0.335
Geometric_Composition	0.411	0.355	0.360
HDR	0.487	0.406	0.475
Hazy	0.493	0.451	0.447
Horror	0.400	0.396	0.295
Long_Exposure	0.515	0.457	0.463
Macro	0.617	0.582	0.530
Melancholy	0.168	0.147	0.136
Minimal	0.512	0.444	0.481
Noir	0.494	0.481	0.408
Pastel	0.258	0.245	0.211
Romantic	0.227	0.204	0.185
Serene	0.281	0.257	0.239
Sunny	0.500	0.481	0.453
Texture	0.265	0.227	0.229
Vintage	0.282	0.273	0.222
mean	0.368	0.336	0.316
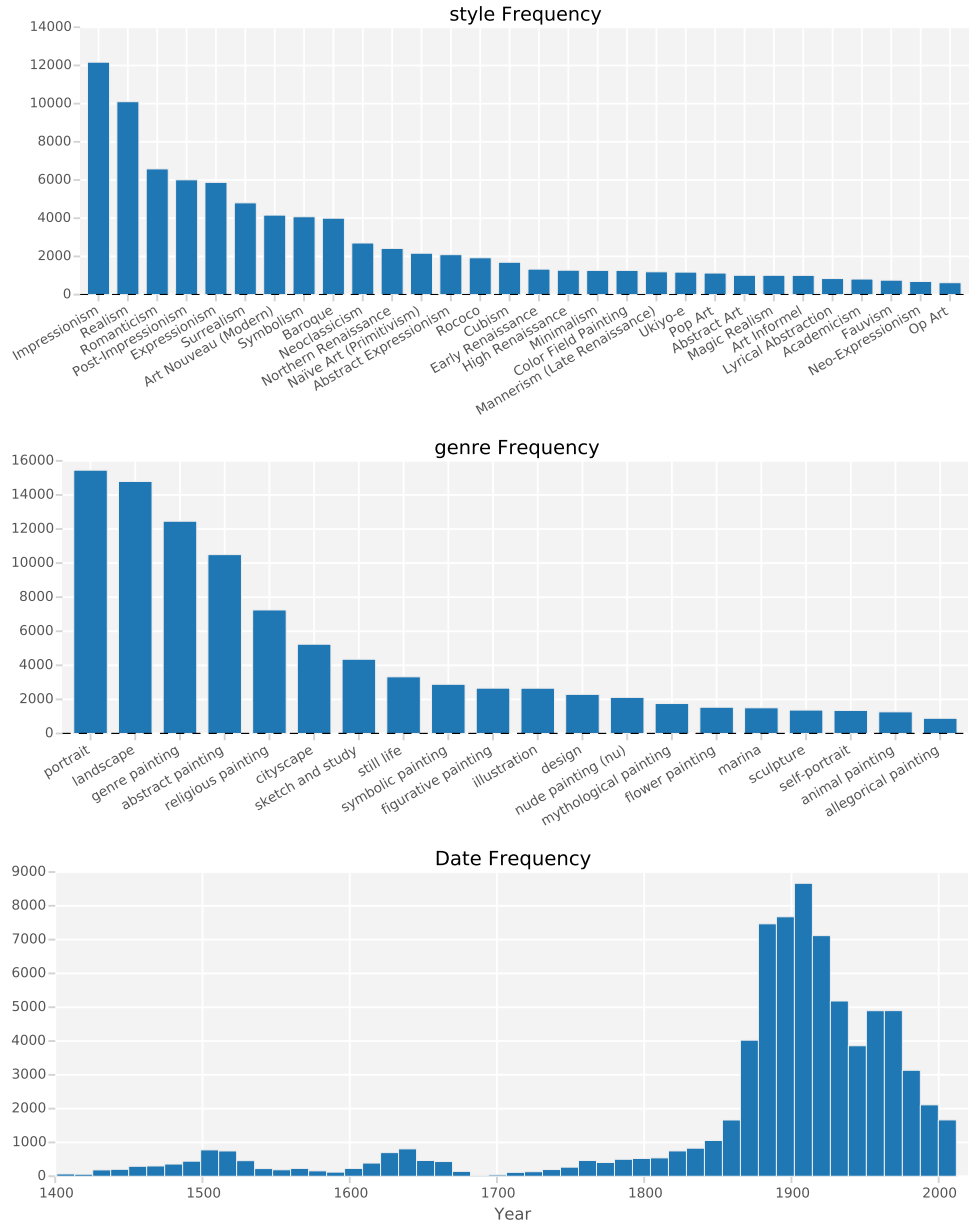
Figure 6: Distribution of image style, genre, and date in the Wikipaintings dataset.

Table 5: All per-class APs on all evaluated features on the Wikipaintings dataset.

	Fusion x Content	MC-bit	DeCAF$_6$
Abstract_Art	0.341	0.314	0.258
Abstract_Expressionism	0.351	0.340	0.243
Art_Informel	0.221	0.217	0.187
Art_Nouveau_(Modern)	0.421	0.402	0.197
Baroque	0.436	0.386	0.313
Color_Field_Painting	0.773	0.739	0.689
Cubism	0.495	0.488	0.400
Early_Renaissance	0.578	0.559	0.453
Expressionism	0.235	0.230	0.186
High_Renaissance	0.401	0.345	0.288
Impressionism	0.586	0.528	0.411
Magic_Realism	0.521	0.465	0.428
Mannerism_(Late_Renaissance)	0.505	0.439	0.356
Minimalism	0.660	0.614	0.604
Nave_Art_(Primitivism)	0.395	0.425	0.225
Neoclassicism	0.601	0.537	0.399
Northern_Renaissance	0.560	0.478	0.433
Pop_Art	0.441	0.398	0.281
Post-Impressionism	0.348	0.348	0.292
Realism	0.408	0.309	0.266
Rococo	0.616	0.548	0.467
Romanticism	0.392	0.389	0.343
Surrealism	0.262	0.247	0.134
Symbolism	0.390	0.390	0.260
Ukiyo-e	0.895	0.894	0.788
mean	0.473	0.441	0.356

Table 6: Comparison of Flickr Style per-class accuracies for our method and Mech Turkers. We first give the full results table, then show the signfcant deviations between human and machine performance, and between using Flickr and MTurk ground truth.

	MTurk acc., Flickr g.t.	Our acc., Flickr g.t.	Our acc., MTurk g.t.
Bright	69.10	73.38	73.63
Depth of Field	68.92	68.50	81.05
Detailed	65.47	75.25	68.44
Ethereal	76.92	80.62	77.95
Geometric Composition	81.52	77.75	80.31
HDR	71.84	82.00	76.96
Hazy	83.49	80.75	81.64
Horror	89.85	84.25	81.64
Long Exposure	73.12	84.19	76.79
Macro	92.25	86.56	88.39
Melancholy	67.77	70.88	71.25
Minimal	79.71	83.75	78.57
Noir	81.35	85.25	85.88
Pastel	66.94	74.56	75.47
Romantic	60.91	68.00	66.25
Serene	69.49	70.44	76.80
Sunny	84.48	84.56	79.94
Vintage	68.77	75.50	67.80
Mean	75.11	78.12	77.15

	Our acc., Flickr g.t.	Our acc., MTurk g.t.	% change from Flickr to MTurk g.t.
Vintage	75.50	67.80	-10.19
Detailed	75.25	68.44	-9.05
Long Exposure	84.19	76.79	-8.79
Minimal	83.75	78.57	-6.18
HDR	82.00	76.96	-6.15
Sunny	84.56	79.94	-5.46
Serene	70.44	76.80	9.03
Depth of Field	68.50	81.05	18.32

	Our acc., Flickr g.t.	MTurk acc., Flickr g.t.	Acc. difference
Horror	84.25	90.42	-6.17
Macro	86.56	91.71	-5.15
Romantic	68.00	61.04	6.96
Pastel	74.56	66.87	7.69
HDR	82.00	72.79	9.21
Long Exposure	84.19	73.83	10.35
Detailed	75.25	63.30	11.95

Table 7: Per-class accuracies on the Wikipaintings dataset, using the MC-bit feature.

Style	Accuracy	Style	Accuracy
Symbolism	71.24	Impressionism	82.15
Expressionism	72.03	Northern Renaissance	82.32
Art Nouveau (Modern)	72.77	High Renaissance	82.90
Nave Art (Primitivism)	72.95	Mannerism (Late Renaissance)	83.04
Surrealism	74.44	Pop Art	83.33
Post-Impressionism	74.51	Early Renaissance	84.69
Romanticism	75.86	Abstract Art	85.10
Realism	75.88	Cubism	86.85
Magic Realism	78.54	Rococo	87.33
Neoclassicism	80.18	Ukiyo-e	93.18
Abstract Expressionism	81.25	Minimalism	94.21
Baroque	81.45	Color Field Painting	95.58
Art Informel	82.09		

Figure 7: Confusion matrix of our best classifier (Late-fusion × Content) on the AVA Style dataset. The right-most "prior" column reflects the distribution of ground-truth labels in the test set. The confusions are mostly understandable: "Soft Focus" vs. "Motion Blur" for example.

Figure 8: Confusion matrix of our best classifier (Late-fusion × Content) on the Flickr dataset.

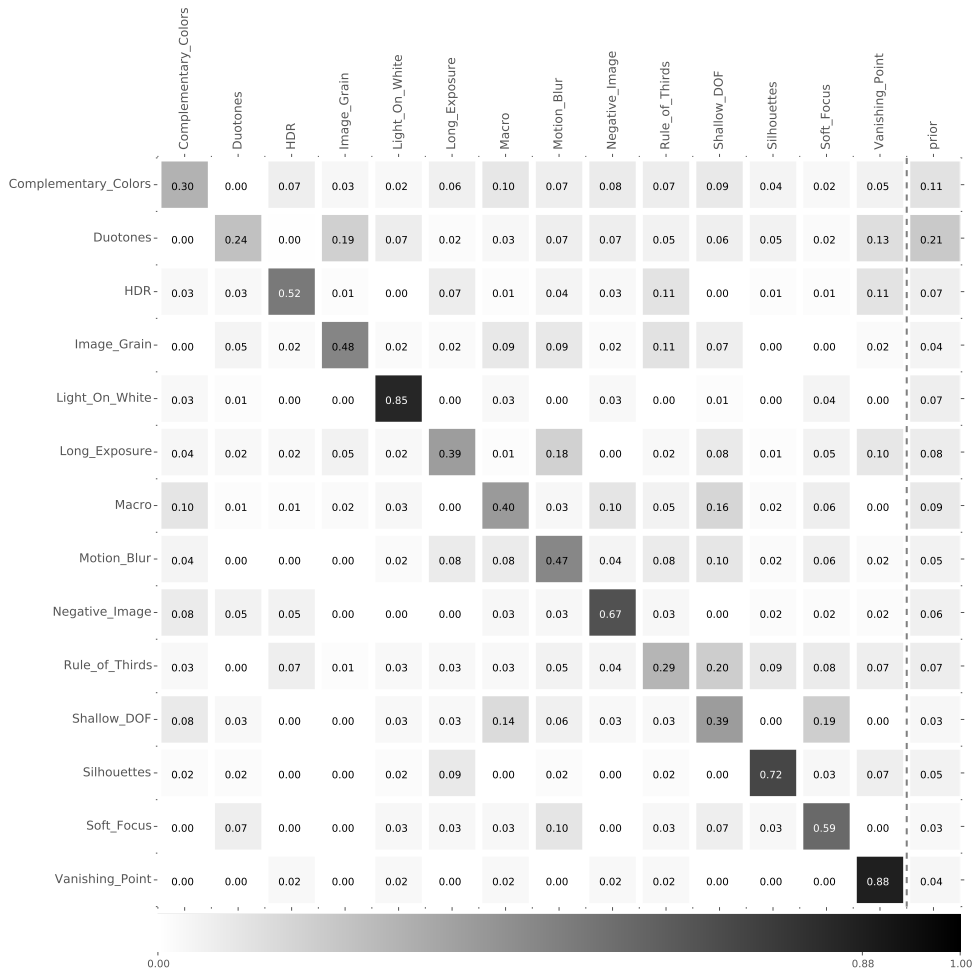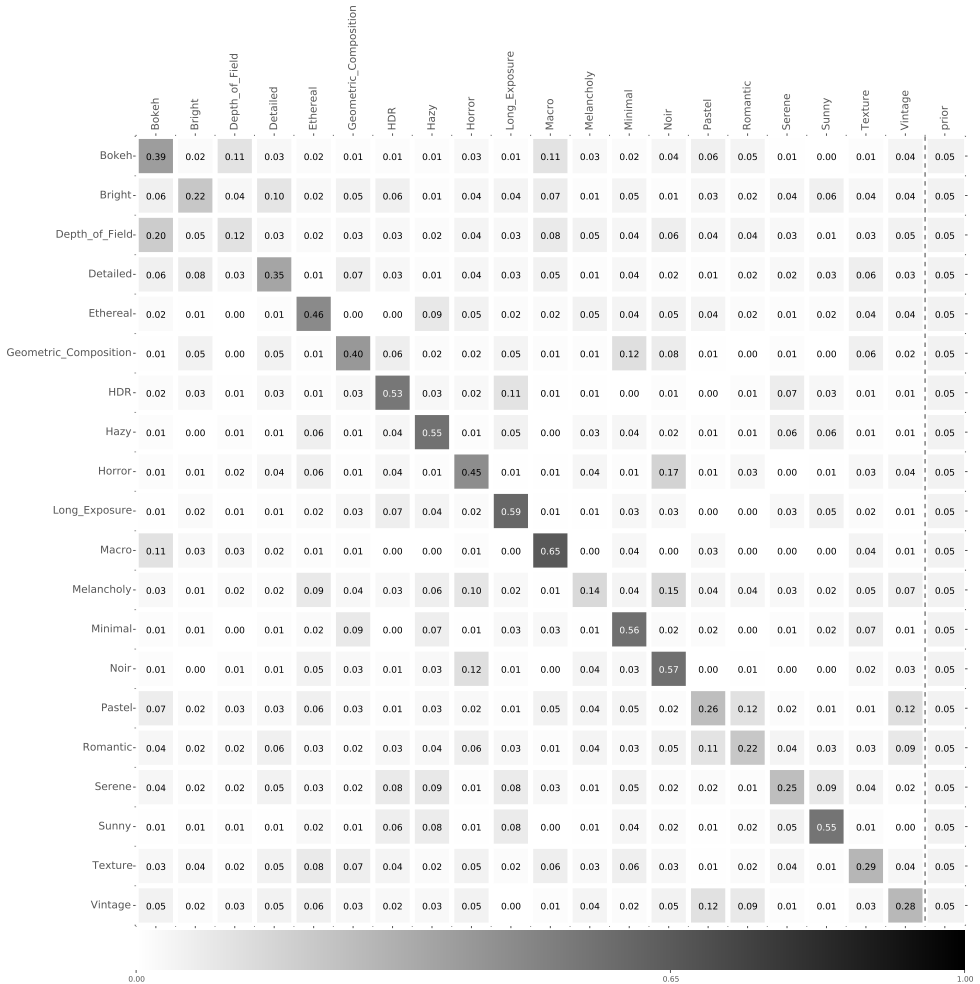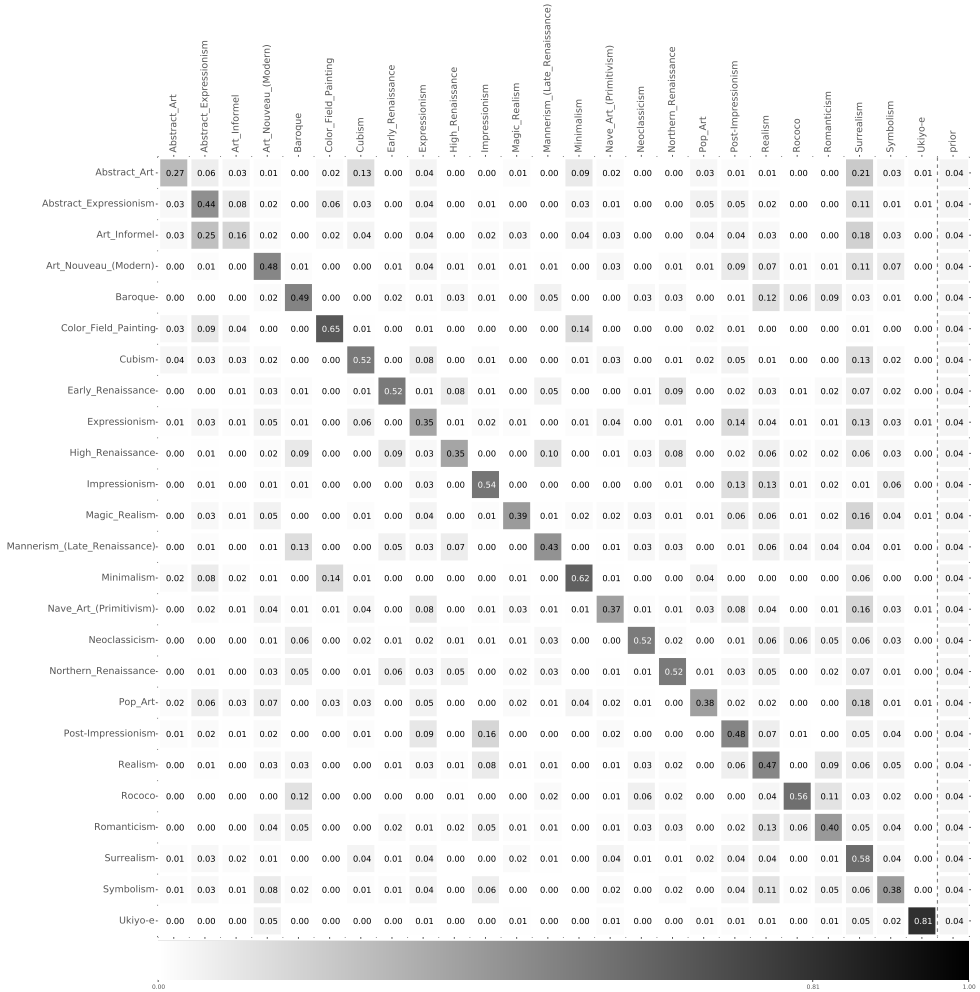|	Abstract_Art	Abstract_Expressionism	Art_Informel	Art_Nouveau_(Modern)	Baroque	Color_Field_Painting	Cubism	Early_Renaissance	Expressionism	High_Renaissance	Impressionism	Magic_Realism	Mannerism_(Late_Renaissance)	Minimalism	Nave_Art_(Primitivism)	Neoclassicism	Northern_Renaissance	Pop_Art	Post-Impressionism	Realism	Rococo	Romanticism	Surrealism	Symbolism	Ukiyo-e	prior
Abstract_Art	0.27	0.06	0.03	0.01	0.00	0.02	0.13	0.00	0.04	0.00	0.00	0.01	0.00	0.09	0.02	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.21	0.03	0.01	0.04
Abstract_Expressionism	0.03	0.44	0.08	0.02	0.00	0.06	0.03	0.00	0.04	0.00	0.01	0.00	0.00	0.03	0.01	0.00	0.00	0.05	0.05	0.02	0.00	0.00	0.11	0.01	0.01	0.04
Art_Informel	0.03	0.25	0.16	0.02	0.00	0.02	0.04	0.00	0.04	0.00	0.02	0.03	0.00	0.04	0.03	0.00	0.00	0.04	0.04	0.03	0.00	0.00	0.18	0.03	0.00	0.04
Art_Nouveau_(Modern)	0.00	0.01	0.00	0.48	0.01	0.00	0.00	0.01	0.04	0.01	0.01	0.01	0.01	0.00	0.03	0.00	0.01	0.01	0.09	0.07	0.01	0.01	0.11	0.07	0.00	0.04
Baroque	0.00	0.00	0.00	0.02	0.49	0.00	0.00	0.02	0.01	0.03	0.01	0.00	0.05	0.00	0.00	0.03	0.03	0.00	0.01	0.12	0.06	0.09	0.03	0.01	0.00	0.04
Color_Field_Painting	0.03	0.09	0.04	0.00	0.00	0.65	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.04
Cubism	0.04	0.03	0.03	0.02	0.00	0.00	0.52	0.00	0.08	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.01	0.02	0.05	0.01	0.00	0.00	0.13	0.02	0.00	0.04
Early_Renaissance	0.00	0.00	0.01	0.03	0.01	0.00	0.01	0.52	0.01	0.08	0.01	0.00	0.05	0.00	0.00	0.01	0.09	0.00	0.02	0.03	0.01	0.02	0.07	0.02	0.00	0.04
Expressionism	0.01	0.03	0.01	0.05	0.01	0.00	0.06	0.00	0.35	0.01	0.02	0.01	0.00	0.01	0.04	0.00	0.01	0.00	0.14	0.04	0.01	0.01	0.13	0.03	0.01	0.04
High_Renaissance	0.00	0.01	0.00	0.02	0.09	0.00	0.00	0.09	0.03	0.35	0.00	0.00	0.10	0.00	0.01	0.03	0.08	0.00	0.02	0.06	0.02	0.02	0.06	0.03	0.00	0.04
Impressionism	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.03	0.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.13	0.13	0.01	0.02	0.01	0.06	0.00	0.04
Magic_Realism	0.00	0.03	0.01	0.05	0.00	0.00	0.01	0.00	0.04	0.00	0.01	0.39	0.01	0.02	0.02	0.03	0.01	0.01	0.06	0.06	0.01	0.02	0.16	0.04	0.01	0.04
Mannerism_(Late_Renaissance)	0.00	0.01	0.00	0.01	0.13	0.00	0.00	0.05	0.03	0.07	0.00	0.00	0.43	0.00	0.01	0.03	0.03	0.00	0.01	0.06	0.04	0.04	0.04	0.01	0.00	0.04
Minimalism	0.02	0.08	0.02	0.01	0.00	0.14	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.62	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.04
Nave_Art_(Primitivism)	0.00	0.02	0.01	0.04	0.01	0.01	0.04	0.00	0.08	0.00	0.01	0.03	0.01	0.01	0.37	0.01	0.01	0.03	0.08	0.04	0.00	0.01	0.16	0.03	0.01	0.04
Neoclassicism	0.00	0.00	0.00	0.01	0.06	0.00	0.02	0.01	0.02	0.01	0.01	0.01	0.03	0.00	0.00	0.52	0.02	0.00	0.01	0.06	0.06	0.05	0.06	0.03	0.00	0.04
Northern_Renaissance	0.00	0.01	0.00	0.03	0.05	0.00	0.01	0.06	0.03	0.05	0.00	0.02	0.03	0.00	0.01	0.01	0.52	0.01	0.03	0.05	0.00	0.02	0.07	0.01	0.00	0.04
Pop_Art	0.02	0.06	0.03	0.07	0.00	0.03	0.03	0.00	0.05	0.00	0.00	0.02	0.01	0.04	0.02	0.01	0.00	0.38	0.02	0.02	0.00	0.00	0.18	0.01	0.01	0.04
Post-Impressionism	0.01	0.02	0.01	0.02	0.00	0.00	0.01	0.00	0.09	0.00	0.16	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.48	0.07	0.01	0.00	0.05	0.04	0.00	0.04
Realism	0.00	0.01	0.00	0.03	0.03	0.00	0.00	0.01	0.03	0.01	0.08	0.01	0.01	0.00	0.01	0.03	0.02	0.00	0.06	0.47	0.00	0.09	0.06	0.05	0.00	0.04
Rococo	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.06	0.02	0.00	0.00	0.04	0.56	0.11	0.03	0.02	0.00	0.04
Romanticism	0.00	0.00	0.00	0.04	0.05	0.00	0.00	0.02	0.01	0.02	0.05	0.01	0.01	0.00	0.01	0.03	0.03	0.00	0.02	0.13	0.06	0.40	0.05	0.04	0.00	0.04
Surrealism	0.01	0.03	0.02	0.01	0.00	0.00	0.04	0.01	0.04	0.00	0.00	0.02	0.01	0.00	0.04	0.01	0.01	0.02	0.04	0.04	0.00	0.01	0.58	0.04	0.00	0.04
Symbolism	0.01	0.03	0.01	0.08	0.02	0.00	0.01	0.01	0.04	0.00	0.06	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.04	0.11	0.02	0.05	0.06	0.38	0.00	0.04
Ukiyo-e	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.05	0.02	0.81	0.04

0.00   0.81   1.00

Figure 9: Confusion matrix of our best classifier (Late-fusion $\times$ Content) on the Wikipaintings dataset.