Recognizing Image Style
Tech Report

Sergey Karayev[1,2] Aaron Hertzmann[3] Holger Winnemöller[3] Aseem Agarwala[3] Trevor Darrel[1,2]
[1] UC Berkeley, [2] ICSI, and [3] Adobe

Abstract

The style of an image plays a significant role in how it is viewed, but 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: 55K Flickr photographs annotated with curated style labels as well as free-form tags, and 85K paintings annotated with style and 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.

Figure 1: Typical images in different style categories of our Flickr Style dataset. The dataset comprises 18 styles in total, each with 3,000 examples.

1. Introduction

Images convey meaning in multiple ways; *visual style* is often a significant component of image meaning for creative images. For example, the same scene portrayed in the lush, springtime colors of a Renoir painting would tell a different story than shown in the harsh, dark tones of a typical horror movie. Visual style is crucial to how a viewer interprets an image in many contexts, including art, design, entertainment, advertising, and social media. Moreover, an increasing amount of visual media consumption though online social media feeds, photo sharing sites, and news sites, is now curated by machines and not people. Yet, virtually no research in computer vision has explored visual style.

This paper introduces new approaches and datasets for the automatic analysis of image style. Visual style is very recognizable to human viewers, yet difficult to define precisely. Style may combine aspects of color, lighting, composition, scene objects, and other facets. Hence, we prefer to define style empirically through labelled data, and then analyze the divisions between these classes. Finding existing datasets insufficient, we gather a new large-scale dataset of photographs annotated with diverse visual style labels. This dataset embodies several different aspects 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”). We also gather a large dataset of visual art (mostly paintings) annotated with art historical style labels, ranging from Renaissance to modern art. We perform a thorough evaluation of different visual features for the task of predicting these style annotations. We find that “deep” features trained on a large amount of data labeled with object class categories (ImageNet) perform significantly better than traditionally used hand-designed features.

The style predictors that our datasets and learning enable are useful as mid-level features in their own right. When making presentations, a searchable source of stylistically

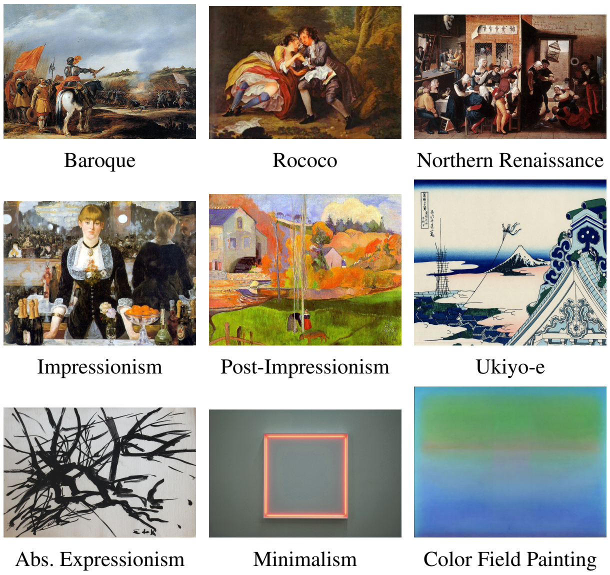

Figure 2: Typical images in different style categories from our Wikipaintings dataset. The dataset comprises 85,000 images labeled with 22 art historical styles.

coherent images would be useful. A story may be illustrated with images that match not only its objective content, but also its sentiment. In addition to evaluating classification performance of our approach, we demonstrate an application of style classifiers to visual search, making a large image collection searchable by both content tags and visual style ("bird, bright/energetic," "train, film noir"). Additionally, we demonstrate that styles learned from paintings can be used to search collections of photographs, and vice versa.

All data, trained predictors, code, and a web-based user interface for searching image collections "with style" will be released upon publication.

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.

The groundwork for predicting aesthetic rating of photographs was laid by Datta et al. [4], who designed visual features to represent concepts such as colorfulness, saturation, rule-of-thirds, and depth of field. Classifiers based on these features were evaluated on a dataset of photographs rated for aesthetics and originality by users of the `photo.net` community. The same approach was further applied to a small set of Impressionist paintings [13].

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]. Their high-level features were themselves trained in a classification framework on labeled datasets. 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.

Murray et al. [17] introduced the Aesthetic Visual Analysis (AVA) dataset, annotated with ratings by users of DPChallenge, a photographic skill competition website. This dataset is primarily aimed at predicting beauty, and Murray et al. showed that generic feature descriptors with state-of-the-art coding gave better predictions than the previously-used hand-designed features. Our use of "deep-network" features trained on a large amount of visual data is informed by their findings.

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. These style labels primarily reflect photographic techniques such as "HDR" and simple compositional qualities like "Duotones." Using images from this dataset, Marchesotti and Peronnin [16] 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.

In contrast to their unsupervised learning approach, we gather annotations of style that are supervised, either by membership in a user-curated Flickr group, or by art historian experts. We are unaware of other previous work gathering annotations of image style.

In a task similar to predicting the style of an image, Borth et al. [3] performed sentiment analysis on images. Following the "ObjectBank" [14] approach, the authors trained and deployed object detectors trained on data labeled with adjective-noun pairs of known sentiment value to predict the sentiment for the entire image.

Features based on image statistics have been successfully employed to detect artistic forgeries, e.g., [15]. Their work focused on extremely fine-scale discrimination between two very similar classes, and has not been applied to broader style classification.

3. Data Sources

Performance of scene and object recognition depends directly on the quality of the training data set. To our knowledge, there is only one existing dataset annotated with visual style, and only a narrow range of styles are represented [17]. We review the best current dataset for aesthetic prediction, which has a subset of style annotations. We then present two new datasets, covering a range of visual styles.

3.1. Aesthetic Visual Analysis (AVA)

AVA [17] is a dataset of 250K images from dpchallenge.net, where users submit and judge photos organized into thematic challenges such as "Cats", "Textures and Materials", or "Depth of Field". On average, an image receives around 200 ratings on a 1-10 scale. These ratings reflect both absolute image quality and how well the image meets the goals of the specific challenge. The dataset also includes 14,000 images annotated with 14 labels of photographic style, manually created by the authors by combining photos from 72 different challenges. The task in this dataset is to predict rating mean and standard deviations, and to predict style labels. The authors did not report prediction scores on style prediction.

The styles in AVA are primarily photographic techniques, such as "HDR" and "Long Exposure," and simple compositional techniques such as "Silhouettes," and "Vanishing Point." Out of the 14 photographic style labels, less than half have over a thousand positive examples, and "higher-level" styles such as mood and genre styles are not represented.

3.2. Flickr Style

Our goal is to describe a broader range of image style beyond photographic style, which also includes a range of genres, compositional styles, and moods. We would like to gather data from a rich source, such as Flickr, so that the size of our dataset can be increased with minimal effort. 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. There are Flickr Groups for most visual style concepts that we considered.

We selected 18 groups with large image collections and clearly defined membership rules to produce our Flickr Style dataset. We collected 3,000 positive examples for each label, for a total of 54,000 images. Example images are shown in Figure 1, and examples of the group rules and image counts are given in Table 1. The names of our styles can be seen in evaluation plots Figure 5 and Figure 6, and in Table 4.

The labels are clean in the positive examples, but 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.

Table 1: Sample of our Flickr Style groups, showing the size of available data and the membership rules.

Group name	Images	Description (excerpt)
Geometric Beauty	153909	Circles, triangles, rectangles, symmetric objects, repeated patterns ...
Pastel, Soft	100116	Pastels and whites, blossom and sunglares. Anything that is soft, pretty and just dreamy.
Film Noir Mood	7775	Not just black and white photography, but a dark, gritty, moody feel. ...
Horror	16501	The scariest pics you can find. ... your bloodiest, ghastliest, freakiest snaps. ...
Melancholy	90748	Only melancholic shots.

3.3. Wikipaintings

The above datasets deal exclusively with modern, photographic images. To our knowledge, no existing vision algorithm can categorize non-photorealistic media, such as paintings and drawings. In a major step to this goal, 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 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. As shown in Figure 3, the dataset presents significant stylistic diversity, primarily spanning Renaissance styles to modern art movements. We select 22 styles with more than 1,000 examples, for a total of 85,000 images. Example images are shown in Figure 2.

4. 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 [15], 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

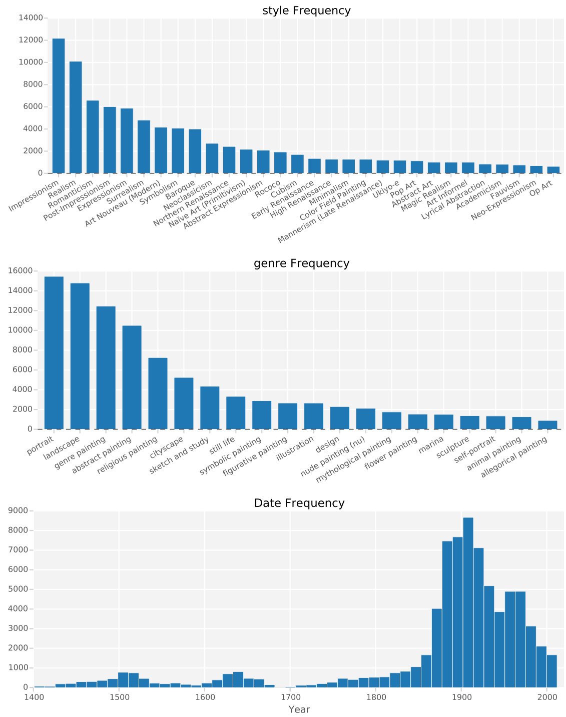

Figure 3: Distribution of image style, genre, and date in the Wikipaintings dataset.

CIELAB color space has 4, 14, and 14 bins in the L*, a*, and b* channels, following Palermo et al. [19], who showed this to be the best performing single feature for determining the date of historical color images.

GIST. The classic gist descriptor [18] 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.

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. We use DeCAF [7], an open-source implementation of such an eight-layer network, trained on over a million images annotated with 1,000 ImageNet classes. We investigate using two different layers of the network, referred to as DeCAF$_5$ (9,000-dimensional) and DeCAF$_6$ (4,000-dimensional, closer to the supervised signal), computed from images center-cropped and resized to 256 by 256 pixels. Since DeCAF is trained on object recognition, not style recognition, we also test whether tuning the network for our style datasets improve performance.

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.

5. Learning algorithm

We wish to learn to classify novel images according to their style, using the labels exemplified by the datasets given 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 enough robustness for linear classification to be accurate.

We use an open-source implementation of Stochastic Gradient Descent with adaptive subgradient [1]. 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. For multi-class classification, we always use the One vs. All reduction. We set the initial learning rate to 0.5, and use adaptive subgradient optimization [8].

For all features except binary ones, values are standardized: each column has its mean subtracted, and is divided by its standard deviation. For feature combinations, we use two-stage late fusion. First, single-feature classifiers are trained, then their scores are linearly combined with weights learned by a second classifier.

6. Evaluation

Details of our experiments follow, with a concluding discussion section.

6.1. AVA Style

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 [17]. Although the provided test split has an uneven class distribution, we found that to compare with the reported results, a class-balanced set is needed.

Consequently, we adhere to the provided split but compute evaluation metrics on a random class-balanced subset of the test data. We use class-balanced data for evaluation in all following experiments.

Metrics. Following the established approach, aesthetic rating is classified in a binary prediction task of being below or above the mean. On this task, we report the accuracy of our predictions.

For multi-class prediction of the style labels, we report the confusion matrix of most confident classifications for each image, top-K accuracies (useful to see when the dataset has easily confused labels), and per-class Average Precision (AP): area under the Precision vs. Recall curve.

Results. For all features, AP scores for the AVA Style dataset are shown in Figure 4. The mean AP scores and the aesthetic rating accuracy are given in the overall results table Table 2.

For aesthetic rating performance, the best single feature is the MC-bit feature, obtaining 0.843 accuracy. Previous work did not report accuracy on this subset of the data, but their best reported accuracy on the test set of the full AVA data is 0.68 [17]. For style classification, the best single feature is the DeCAF$_6$ convolution network feature, obtaining 0.577 mean AP. Feature fusion improves the result to 0.604

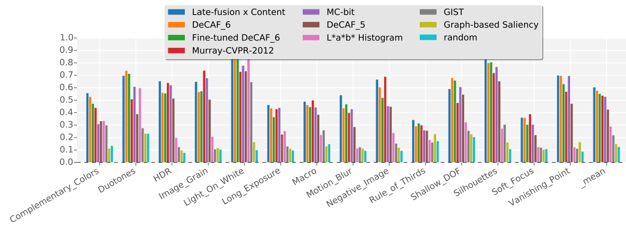

Figure 4: APs on the AVA Style dataset.

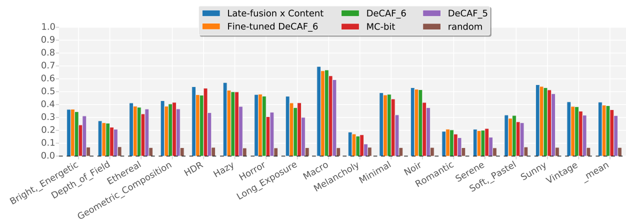

Figure 5: APs on the Flickr dataset.

mean AP; both results beat the previous state-of-the-art of 0.538 mean AP [17].

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

6.2. Flickr Style

Following the same evaluation setup and metrics as above, we learn and predict style labels on the 53,000 images labeled with 18 different visual styles of our new Flickr Style dataset, using 20% of the data for testing, and another 20% for parameter tuning validation. Results are presented in Figures 5 and 9, and in Table 2.

The best single-channel feature is again DeCAF$_6$ with 0.396 mean AP; feature fusion obtains 0.419 mean AP. Surprisingly, fine-tuning the DeCAF convolution net with images from our datasets did not increase performance.

Content correlations. We plot the confusion matrix of this single-label dataset in Figure 6. As expected, there are 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 7. We can observe that some styles have strong correlations to content (e.g. "Hazy"

Table 2: Mean APs (or accuracies, where noted) on three datasets for the considered single-channel features and their second-stage combination. As some features were clearly dominated by others on the AVA dataset, only the better features were evaluated on larger datasets.

	Late-fusion	DeCAF$_5$	DeCAF$_6$	MC-bit	Tuned DC$_6$	L*a*b* Hist	GIST	Saliency	random
AVA Rating (acc.)	-	0.779	0.686	0.843	0.720	0.574	0.558	0.539	0.500
AVA Style	0.604	0.427	0.577	0.529	0.552	0.291	0.220	0.149	0.127
Flickr	0.419	0.314	0.391	0.360	0.396	-	-	-	0.066
Wikipaintings	0.476	-	0.356	0.443	0.356	-	-	-	0.043

occurs with "vehicle", "HDR" doesn't occur with "cat"). To further enable our linear classifier to take advantage of such correlations, we take an outer product of our content classifier features with the second-stage late fusion features ("Late-fusion × Content" in all results figures).

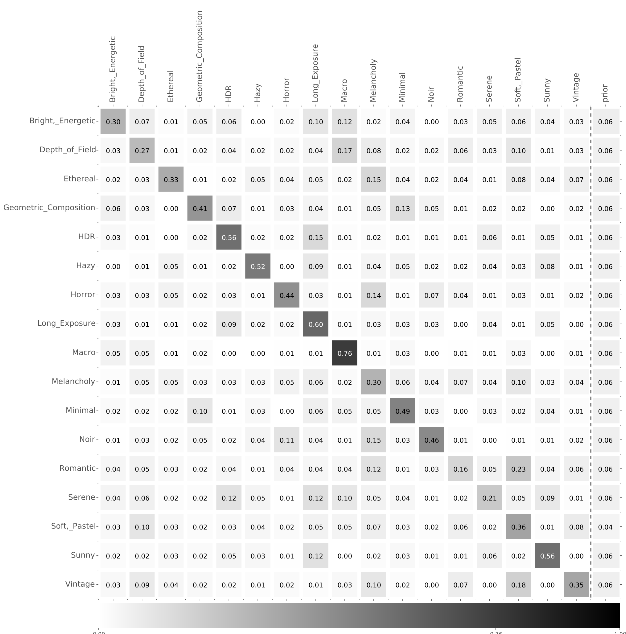

Figure 6: Confusion matrix of our best classifier (Late-fusion × Content) on the Flickr dataset.

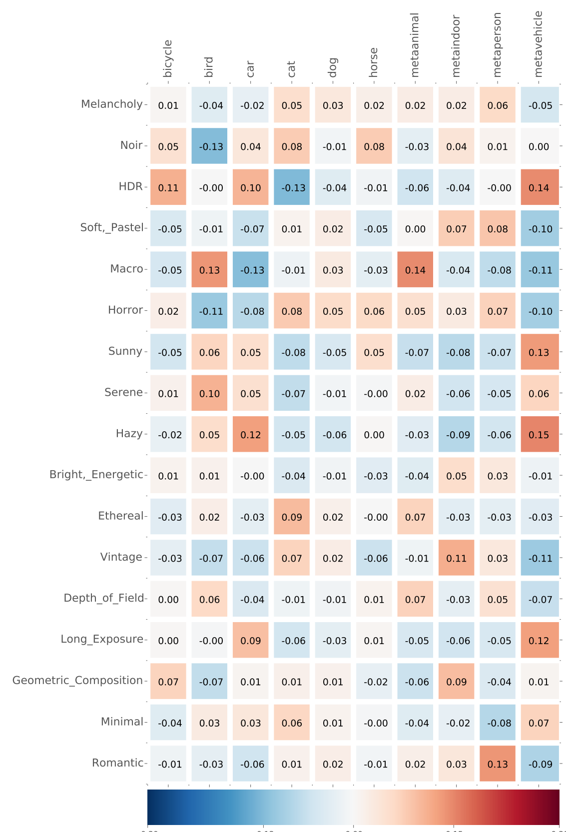

Figure 7: Correlation of PASCAL content classifers (columns) against ground truth Flickr style labels (rows).

6.3. Wikipaintings

With the same setup and features as in the Flickr experiments, we evaluate 85,000 images labeled with 22 different art styles. The results are given in Figures 8 and 9, and in Table 2. The best single-channel feature is MC-bit with 0.444 mean AP; feature fusion obtains 0.476 mean AP. As with Flickr, fine-tuning the convolutional net feature did not increase its performance on this dataset.

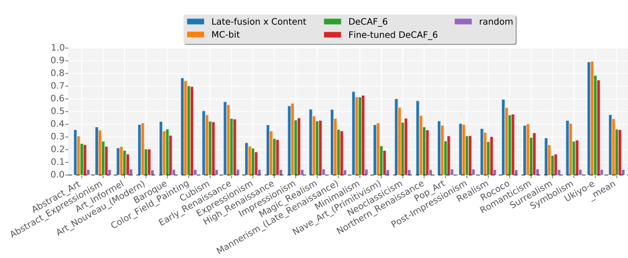

Figure 8: APs on the Wikipaintings dataset.

6.4. Applications of style classifiers

Style classifiers learned on our datasets can be used toward novel goals. For example, sources of stock photog-

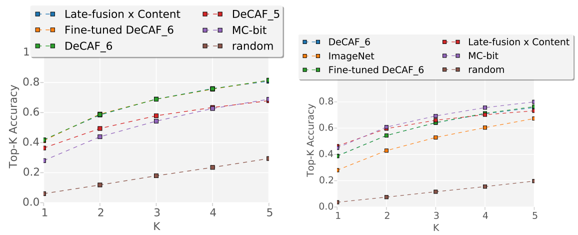

Figure 9: Top-K accuracies for the Flickr and Wikipaintings datasets, respectively.

raphy 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. We apply style classifiers to the PASCAL visual object class dataset, and show top hits for different styles for the "bird" and "train" categories in Figure 11.

Additionally, 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 10.

7. Conclusion

We have described datasets and algorithms for classifying image styles. Given the importance of style in modern visual communication, we believe that understanding style is an important challenge for computer vision, and our results illustrate the potential for future research in this area.

One challenging question is to define and understand the meaning of style. Different types of styles relate to content, color, lighting, composition, and other factors. Our work provides some preliminary evidence about the relationships of these quantities.

We were suprised by the success of the DeCAF convolution net, which was originally trained for object recognition. Moreover, fine-tuning it for style did not significantly increase performance. 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 recogntion depends on object appearance, e.g., distinguishing red from white wine, or different kinds of terriers, and that the model learns to repurpose these feature for image style.

Another possibility is that the style labels can be predicted from object content alone. We do see strong correlations in our data, e.g., *Macro* images frequently depict birds and flowers.However, we found that using 1,000 ImageNet classifiers as a features was significantly worse than the performance of the DeCAF_6 layer feature (see Tables 3, 4, 5).

References

[1] A. Agarwal, O. Chapelle, M. Dudik, and J. Langford. A Reliable Effective Terascale Linear Learning System. *Journal of Machine Learning Research*, 2012. 4

[2] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*, 2012. 4

[3] D. Borth, R. Ji, T. Chen, and T. M. Breuel. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM MM*, 2013. 2

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*, 2006. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3

[6] S. Dhar, T. L. Berg, and S. Brook. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR*, 2011. 2, 4

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, T. Eecs, and B. Edu. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. Technical report, 2013. arXiv:1310.1531v1. 4

[8] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 2011. 5

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL VOC Challenge Results, 2010. 4

[10] M. Gygli, F. Nater, and L. V. Gool. The Interestingness of Images. In *ICCV*, 2013. 2

[11] J. Harel, C. Koch, and P. Perona. Graph-Based Visual Saliency. In *NIPS*, 2006. 4

[12] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR*, June 2011. 2

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

[14] L.-j. Li, H. Su, E. P. Xing, and L. Fei-fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *NIPS*, 2010. 2

[15] S. Lyu, D. Rockmore, and H. Farid. A digital technique for art authentication. *PNAS*, 101(49), 2004. 2, 3

[16] L. Marchesotti and F. Perronnin. Learning beautiful (and ugly) attributes. In *BMVC*, 2013. 2

[17] N. Murray, D. Barcelona, L. Marchesotti, and F. Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, 2012. 2, 3, 5

[18] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42(3):145–175, 2001. 4

[19] F. Palermo, J. Hays, and A. A. Efros. Dating Historical Color Images. In *ECCV*, 2012. 3

Figure 10: 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.)

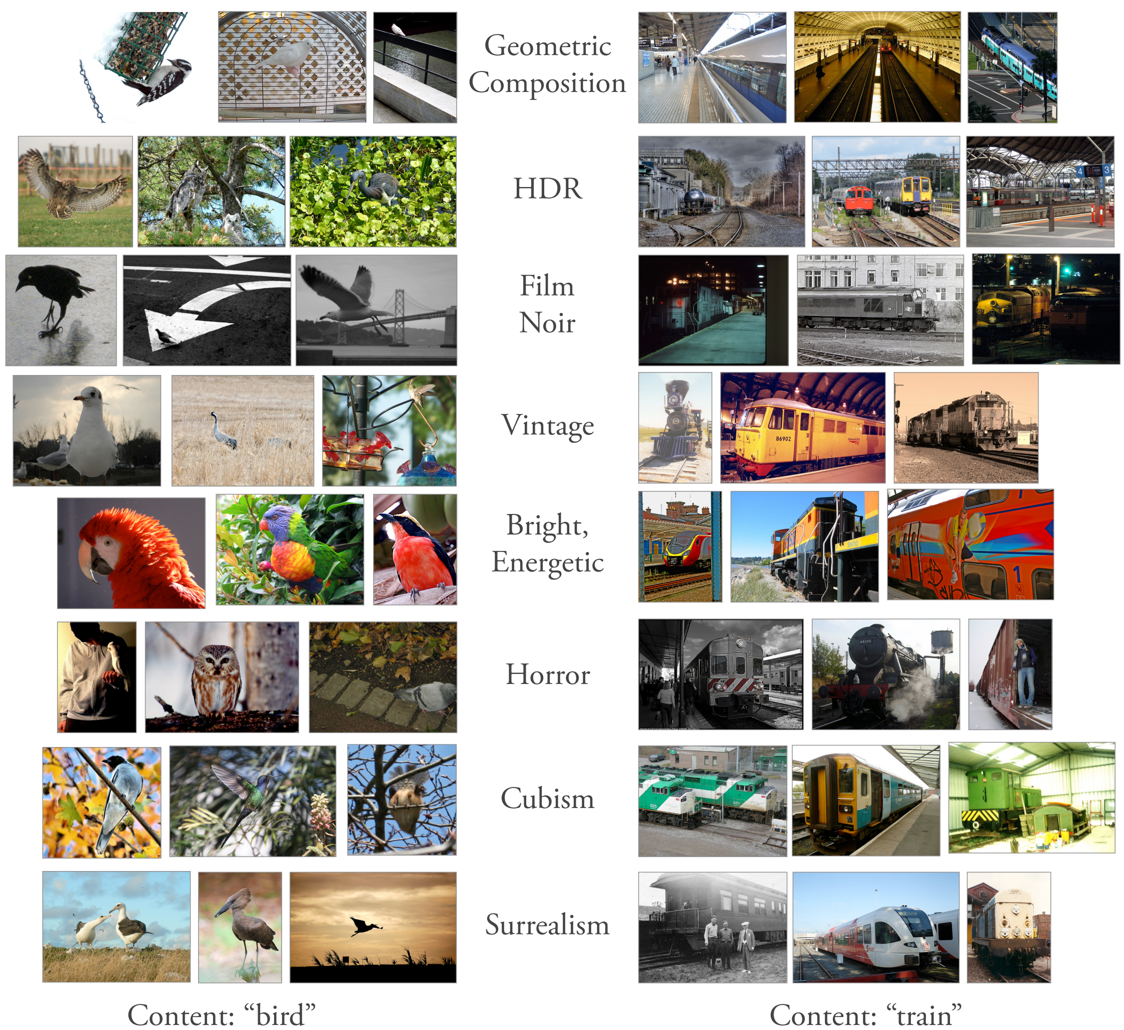

Figure 11: Style-based search within the PASCAL dataset, showing the top 3 images with classifier "bird" and "train," in six Flickr styles and two Wikipaintings style. Note that the PASCAL data includes only 773 bird images and 550 trains, and thus their stylistic range is limited. (Figure best viewed in color.)

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

	Late-fusion	$DeCAF_6$	$DC_{6,ft}$	MC-bit	Murray	$DeCAF_5$	ImageNet	L*a*b*	GIST	Saliency
Complementary_Colors	0.469	0.548	0.514	0.329	0.440	0.368	0.389	0.294	0.223	0.111
Duotones	0.676	0.737	0.665	0.612	0.510	0.363	0.383	0.582	0.255	0.233
HDR	0.669	0.594	0.516	0.624	0.640	0.494	0.335	0.194	0.124	0.101
Image_Grain	0.647	0.545	0.563	0.744	0.740	0.535	0.219	0.213	0.104	0.104
Light_On_White	0.908	0.915	0.860	0.802	0.730	0.805	0.508	0.867	0.704	0.172
Long_Exposure	0.453	0.431	0.444	0.420	0.430	0.208	0.242	0.232	0.159	0.147
Macro	0.478	0.427	0.488	0.413	0.500	0.376	0.438	0.230	0.269	0.161
Motion_Blur	0.478	0.467	0.380	0.458	0.400	0.327	0.186	0.117	0.114	0.122
Negative_Image	0.595	0.619	0.561	0.499	0.690	0.427	0.323	0.268	0.189	0.123
Rule_of_Thirds	0.352	0.353	0.290	0.236	0.300	0.269	0.244	0.188	0.167	0.228
Shallow_DOF	0.624	0.659	0.627	0.637	0.480	0.522	0.517	0.332	0.276	0.223
Silhouettes	0.791	0.801	0.835	0.801	0.720	0.609	0.401	0.261	0.263	0.130
Soft_Focus	0.312	0.354	0.305	0.290	0.390	0.225	0.170	0.127	0.126	0.114
Vanishing_Point	0.684	0.658	0.646	0.685	0.570	0.527	0.542	0.123	0.107	0.161
mean	0.581	0.579	0.550	0.539	0.539	0.432	0.350	0.288	0.220	0.152

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

	Late-fusion x Content	Fine-tuned $DeCAF_6$	$DeCAF_6$	MC-bit	$DeCAF_5$	Imagenet
Bright,_Energetic	0.355	0.340	0.331	0.250	0.313	0.231
Depth_of_Field	0.266	0.252	0.241	0.230	0.208	0.202
Ethereal	0.418	0.383	0.365	0.328	0.356	0.190
Geometric_Composition	0.442	0.409	0.395	0.399	0.369	0.347
HDR	0.548	0.488	0.477	0.527	0.332	0.293
Hazy	0.565	0.504	0.506	0.489	0.386	0.330
Horror	0.479	0.471	0.464	0.304	0.337	0.286
Long_Exposure	0.469	0.415	0.388	0.426	0.300	0.254
Macro	0.684	0.667	0.683	0.620	0.588	0.640
Melancholy	0.178	0.166	0.157	0.169	0.096	0.131
Minimal	0.498	0.476	0.465	0.452	0.319	0.281
Noir	0.529	0.527	0.521	0.409	0.372	0.290
Romantic	0.200	0.210	0.206	0.162	0.140	0.185
Serene	0.209	0.197	0.191	0.219	0.142	0.175
Soft,_Pastel	0.309	0.304	0.317	0.267	0.269	0.272
Sunny	0.550	0.551	0.540	0.523	0.481	0.388
Vintage	0.421	0.382	0.385	0.348	0.309	0.268
mean	0.419	0.397	0.390	0.360	0.313	0.280

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

	Late-fusion x Content	MC-bit	$DeCAF_6$	Fine-tuned $DeCAF_6$	ImageNet
Abstract_Art	0.341	0.314	0.258	0.233	0.192
Abstract_Expressionism	0.351	0.340	0.243	0.222	0.159
Art_Informel	0.221	0.217	0.187	0.158	0.138
Art_Nouveau_(Modern)	0.421	0.402	0.197	0.219	0.096
Baroque	0.436	0.386	0.313	0.330	0.162
Color_Field_Painting	0.773	0.739	0.689	0.703	0.503
Cubism	0.495	0.488	0.400	0.427	0.193
Early_Renaissance	0.578	0.559	0.453	0.424	0.192
Expressionism	0.235	0.230	0.186	0.186	0.093
High_Renaissance	0.401	0.345	0.288	0.281	0.165
Impressionism	0.586	0.528	0.411	0.433	0.227
Magic_Realism	0.521	0.465	0.428	0.435	0.198
Mannerism_(Late_Renaissance)	0.505	0.439	0.356	0.359	0.171
Minimalism	0.660	0.614	0.604	0.636	0.449
Nave_Art_(Primitivism)	0.395	0.425	0.225	0.210	0.111
Neoclassicism	0.601	0.537	0.399	0.438	0.179
Northern_Renaissance	0.560	0.478	0.433	0.339	0.119
Pop_Art	0.441	0.398	0.281	0.304	0.163
Post-Impressionism	0.348	0.348	0.292	0.317	0.135
Realism	0.408	0.309	0.266	0.265	0.159
Rococo	0.616	0.548	0.467	0.501	0.242
Romanticism	0.392	0.389	0.343	0.265	0.185
Surrealism	0.262	0.247	0.134	0.152	0.099
Symbolism	0.390	0.390	0.260	0.296	0.172
Ukiyo-e	0.895	0.894	0.788	0.765	0.260
mean	0.473	0.441	0.356	0.356	0.191